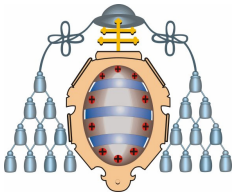
UNIVERSIDAD DE OVIEDO
Departamento de Informática



TESIS DOCTORAL

blindLight
Una nueva técnica para procesamiento de texto no estructurado mediante vectores de *n*-gramas de longitud variable con aplicación a diversas tareas de tratamiento de lenguaje natural

Presentada por

Daniel Gayo Avello

para la obtención del título de Doctor por la Universidad de Oviedo

Dirigida por el

Profesor Doctor D. Darío Álvarez Gutiérrez

Oviedo, Junio de 2005

RESUMEN

Es posible transformar, de manera automática, textos de cualquier idioma alfabético en vectores de n-gramas de longitud variable capaces de almacenar ciertos aspectos de la semántica subyacente al texto inicial. Estos vectores pueden transformar la información original, ser comparados e incluso combinados entre sí subrayando, como resultado, gran parte de la semántica presente en el texto de partida.

Se han utilizado frecuentemente n-gramas para llevar a cabo distintas tareas de tratamiento de lenguaje natural. La mayoría de estas técnicas tienen algunos puntos en común: (1) los documentos son mapeados sobre un espacio vectorial donde los n-gramas son utilizados como coordenadas y las frecuencias relativas de aparición de los mismos en el texto como pesos del vector, (2) muchas de estas técnicas producen un contexto para cada documento que juega un papel similar al de las listas de "palabras vacías" (stop-words) y (3) el coseno del ángulo formado por los vectores de los documentos se utiliza normalmente para determinar la similitud entre documentos o entre consultas y documentos.

blindLight es una nueva propuesta, desarrollada por este doctorando, relacionada con tales técnicas "clásicas" aunque introduce dos importantes diferencias: (1) no se utilizan las frecuencias relativas como pesos de los vectores sino las significatividades de los n-gramas y (2) se descarta el coseno del ángulo entre vectores de documentos en favor de una nueva métrica inspirada por las técnicas de alineación de secuencias aunque no tan costosa computacionalmente.

Esta nueva propuesta puede ser utilizada simultáneamente para categorizar u obtener grupos de documentos, recuperar información o extraer frases clave y resúmenes a partir de un único documento. Muchas de estas tareas son herramientas fundamentales para aliviar la "sobrecarga de información" y mejorar la experiencia de los usuarios.

ABSTRACT

It is possible to automatically transform texts written in any western language in variable-length n-gram vectors which preserve some of the semantics from the source texts. Such vectors can transform the primary information, be compared and even combined with each other highlighting, as a result, much of the semantics from the original document.

N-grams have been frequently used to perform different natural language processing tasks. Such methods show many features in common: (1) documents are represented using a vector space where n-grams are taken as coordinates and n-gram frequencies within documents as vector weights, (2) many of these techniques require a background which plays a role similar to that of lists of stop words and (3) the cosine similarity is normally used to compare documents to each other and documents to queries.

blindLight is a new approach, proposed by this researcher, related to such "classical" methods but with two major changes: (1) n-gram relative frequencies within documents are no more used as vector weights but their significances and (2) cosine distance is abandoned in favor of a new measure inspired by sequence alignment techniques although not so computationally expensive.

Such a new proposal can be used to perform automatic document clustering and categorization, information retrieval, in addition to keyphrase extraction and automatic summarization. Such tasks are essential tools to fight "information overload" and improve user experience.

AGRADECIMIENTOS

Mamá, va por ti.

Si he aprendido algo del largo proceso que me ha llevado hasta la conclusión de este trabajo han sido dos cosas: la investigación requiere rigurosidad y, sobre todo, humildad. He hecho cuanto estuvo en mi mano para conseguir lo primero y debo decir que espero de corazón que mi futuro trabajo sea mejor que esta limitada contribución.

Por otro lado, es sabido que un trabajo de esta índole nunca es posible en solitario (como atestiguan dieciocho páginas de referencias) y muchas personas han contribuido a que yo pudiera llevarlo a cabo. Mencionar a todos y cada uno de los que, de un modo u otro, han participado en mi formación académica y profesional resultaría demasiado extenso y podría olvidarme de alguien, así que lo haré de manera breve y sin entrar en detalles, cada uno de ellos sabe a qué me refiero.

En primer lugar me siento agradecido a los hombres y mujeres que forman el Departamento de Informática de la Universidad de Oviedo, desde la dirección al PAS, pasando por compañeros y alumnos. Todos ellos hacen que éste sea un lugar en el que, a pesar de las adversidades, siga siendo una satisfacción y un orgullo trabajar. Un recuerdo especial para Brugos, Cobas y María Jesús que tanto me ha ayudado con el papeleo.

Naturalmente, en un departamento tan grande es imposible tener la misma familiaridad con todo el mundo y si hay un grupo donde me he sentido acogido (y protegido) es Oviedo3, un abrazo para Cueva y Benjamín y un beso para Marián y Almudena, sin ellas Introducción a la Programación me habría sobrepasado.

No puedo dejar de recordar a los compañeros con los que compartí durante cuatro cursos "El Palomar". La incomodidad, el calor y la falta de espacio se olvidaban gracias a vosotros. De entre ellos debo citar a los dos caballeros que durante ese tiempo formaron conmigo un 3 en raya humano: Guti y Luis. Cada conversación que tuvimos fue un placer y mucho de lo que hablamos me ayudó a terminar este trabajo o me confortó en momentos difíciles. A Luis debo agradecerle tantos correos inspiradores y aquella traducción al sueco. A Guti he de agradecerle muchas cosas pero sólo mencionaré una: su libro acerca de cómo escribir una tesis, debes terminarlo.

Mención aparte merece Darío, director de esta tesis, que me ha llevado casi de la mano por los difíciles caminos de la investigación y que mucho antes demostró una confianza en mí por la que siempre le estaré agradecido.

Pero por encima de todo tengo una deuda infinita con dos mujeres que ahora mismo se sienten muy aliviadas: Tensi y mi madre. Ambas ocupan ex aequo mi corazón y no creo que exagere si digo que ellas han sufrido con esta tesis más que yo y que, en consecuencia, es más un logro suyo que mío. Por todo vuestro apoyo, amor y, sobre todo, paciencia, ¡gracias! Aunque hubiese sido capaz de terminar sin esas tres cosas no habría merecido la pena.

TABLA DE CONTENIDOS

CAPÍTULO 1

INTRODUCCIÓN

Hace aproximadamente cincuenta años que se comenzaron a utilizar ordenadores para almacenar y tratar información textual. Se esperaba que esta automatización pondría el conocimiento al alcance de todos e impulsaría enormemente el avance científico puesto que la información estaría, literalmente, en la punta de los dedos de los usuarios. La aparición primero de Internet y más tarde de la Web parecían acercar aún más ese ideal de una "biblioteca universal". Desafortunadamente, las enormes capacidades de almacenamiento de los soportes informáticos y el talento de la especie humana para producir nueva información a un ritmo increíble han hecho de la Web no sólo la biblioteca más grande que haya existido hasta el momento sino también la más anárquica. Se han investigado decenas de propuestas para llevar algo de orden a la Web. La última de ellas pretende que ésta dé un paso más en su evolución hasta convertirse en una Web Semántica permitiendo que agentes software sean capaces de localizar datos, realizar razonamientos y construir nuevo conocimiento de manera autónoma. No obstante, no parece que el texto no estructurado vaya a desaparecer de la Web, es más, probablemente no deje de aumentar. Por esa razón será necesario ofrecer mecanismos complementarios a la Web Semántica a fin de ayudar a los usuarios a lidiar con esa otra "Web no-tan-Semántica". El autor ha propuesto a ese fin la Web Cooperativa que precisará el uso de técnicas ya existentes y la investigación de algunas nuevas. Una de esas últimas es una técnica desarrollada por el autor, sencilla, independiente del idioma y susceptible de ser aplicada a diversas tareas de tratamiento de lenguaje natural. La motivación, fundamentos y viabilidad de dicha técnica así como su eventual aplicación al problema de la sobrecarga de información son objeto de estudio en este trabajo.

1 Almacenamiento y tratamiento automatizado de información

Los primeros ordenadores electrónicos eran poco más que calculadoras no requiriendo dispositivos de almacenamiento externo demasiado sofisticados. Bastaba, tan sólo, disponer de algún método para conservar el código y quizás también los datos de tal forma que no fuese necesario introducirlos "manualmente" cada vez que se desease ejecutar el correspondiente programa.

Cintas y tarjetas perforadas servían perfectamente a ese objetivo, siendo las últimas las más utilizadas en las máquinas comerciales aun cuando la cantidad de información que se podía codificar en cada una fuese limitada. El formato de *IBM*, por ejemplo, permitía almacenar en el denominado "modo texto" 80 caracteres por tarjeta. Así, un bloque de tarjetas de una pulgada de grosor, que contenía unas 143, almacenaría 11.440 caracteres, o lo que es lo mismo, un texto de unas 1.800 palabras. Dicho de otro modo, utilizando semejante soporte una comunicación a un congreso "típica", alrededor de 5.000 palabras, ocuparía 18,73 x 8,26 x 7,06cm, algo más de mil centímetros cúbicos.

Así pues, almacenar grandes cantidades de texto en tarjetas perforadas, aunque posible, resultaba indudablemente incómodo. La Seguridad Social de los EE.UU. debió experimentar semejantes problemas al acercarse la década de los 50 cuando mantenía toda la información sobre los trabajadores del país en tarjetas perforadas puesto que, aparentemente, presionó a *IBM* (2002) para solucionar esta situación; lo cual llevó a la empresa a establecer en 1952 el estándar *de facto* para almacenamiento en cinta magnética (aunque el *UNIVAC I* ya había utilizado ese soporte en 1951).

No obstante, a pesar de sus ventajas sobre las tarjetas perforadas (menor volumen, mayor velocidad de acceso, o posibilidad de reescritura de datos), las cintas presentaban el problema del acceso secuencial. En 1956 *IBM* presentó el *RAMAC 305* (*Random Access Method of Accounting and Control*, Método de Acceso Aleatorio para Contabilidad y Control) que incluía una unidad de almacenamiento en disco magnético, el *IBM 350*, con acceso aleatorio y capacidad para 5 millones de caracteres (o lo que es lo mismo, 62.500 tarjetas perforadas, 2 rollos de cinta para la unidad *IBM 726* de 1952, o 750.000 palabras).

El desarrollo de ambos tipos de soporte continuó, aumentando de manera progresiva tanto la densidad de almacenamiento como la velocidad de acceso, relegando a las tarjetas perforadas a tareas de introducción de datos hasta su práctica desaparición a mediados de los años 70. En aquel momento un rollo de cinta podía almacenar 180MB de datos y la unidad de disco *IBM 3340* 70MB (el equivalente a 230.000 y 90.000 tarjetas perforadas, respectivamente).

Por otro lado, aunque esta época estuvo caracterizada, desde el punto de vista del tratamiento de datos, por el uso de *mainframes* para almacenar y procesar básicamente registros y transacciones, esto es, información estructurada, existía también una cantidad enorme de información textual con poca o ninguna estructura que crecía de un modo continuo y debía ser consultada con frecuencia (patentes, jurisprudencia, informes técnicos, memorandos, etc.)

A partir de mediados de los 50 y en especial en los años 60 se implementaron múltiples sistemas de búsqueda en distintas organizaciones. Según Madeline M. Henderson (1998) existían en 1966 sólo en EE.UU. más de 150 sistemas automatizados para la consulta de información textual. No es de extrañar pues que algunos de los trabajos más influyentes en el campo del tratamiento y recuperación de información surgieran precisamente en esta época. No obstante, hacer una revisión exhaustiva de los primeros años de investigación en dicho área va más allá del objetivo de este trabajo. Para proseguir la línea argumental bastará con exponer algunos hitos fundamentales; el lector interesado en el tema puede consultar el interesante trabajo de Mary Elizabeth Stevens (1970).

Puede considerarse a Hans Peter Luhn el pionero del área. Describió un método estadístico para codificar y, posteriormente, recuperar información textual[1] de forma totalmente automática (Luhn 1957) y una técnica para obtener resúmenes automáticos (Luhn 1958). Luhn proponía utilizar la frecuencia de aparición en el texto de las distintas palabras, obviando las poco frecuentes y las demasiado comunes, introduciendo así el uso de la frecuencia de los términos en cada documento y de listas de ***stop words*** ("palabras vacías[2]"). Ambas técnicas siguen en uso.

Poco después, Melvin E. Maron y J.L. Kuhns (1960) propusieron una alternativa aritmética a la búsqueda booleana (los términos de la **consulta** están o no presentes en los documentos) que permitiría calcular para cada documento una cifra que indicase su mayor o menor grado de **relevancia**[3] en relación con la consulta planteada. Ellos son los primeros en señalar que la ponderación de los distintos términos tanto en la consulta como en los documentos de la **colección** es fundamental y que es posible asignar un simple "número" a un documento para indicar su mayor o menor relevancia para una consulta dada.

Aparentemente, los "pesos" de cada término debían ser asignados manualmente, por el usuario en el caso de las consultas y por un "bibliotecario" en el de los documentos. No obstante, señalan que la relevancia de los términos con que se etiqueta un documento será inversamente proporcional al número de documentos etiquetados, algo que, posteriormente, se revelaría muy importante.

El modelo propuesto presentaba otra dificultad más: para cada documento de la colección, D_i, y cada posible término empleado en una consulta, t_j, se debe conocer la probabilidad de que un usuario en busca de información del tipo contenido en el documento D_i emplease el término t_j en su consulta. Así pues, aun cuando en el experimento descrito por Maron y Kuhns se utilizaron palabras clave extraídas de los propios documentos, en colecciones realmente grandes este proceso sería muy difícil. No obstante, y a pesar de estos inconvenientes de índole práctica, la importancia de las ideas planteadas en ese trabajo es indudable.

Desde mediados de los 60 Gerald Salton y su equipo desarrollaron el sistema de recuperación de información *SMART* (*System for the Mechanical Analysis and Retrieval of Text*, Sistema para el Análisis y Recuperación Mecánica de Texto) introduciendo toda una serie de conceptos de gran influencia posterior: el modelo vectorial de documentos, la utilización de la función coseno para comparar consultas con documentos (Salton y Lesk 1965),

---

[1] El problema básico de la recuperación de información textual consiste en la forma de representar un conjunto (o colección) de documentos no estructurados (texto libre) para facilitar posteriormente la localización de aquellos que satisfagan una necesidad de información de un usuario formulada mediante una consulta también textual.

[2] Aquellas palabras que, a pesar de un uso frecuente, aportan por sí solas poco significado a un texto (se muestran subrayadas algunas palabras vacías del castellano).

[3] La relevancia es una medida de la "proximidad" entre los contenidos de un documento y la necesidad de información planteada por un usuario en forma de consulta. Está claro que se tratará no sólo de un valor subjetivo sino también cambiante, por lo que el término no suele hacer referencia al "juicio" que emitiría un usuario sino al valor que un sistema automático asigna a cada documento en relación con una consulta. El objetivo de los sistemas de recuperación de información es producir valores de relevancia próximos a los que asignaría el propio usuario.

algoritmos de ***stemming***[1] o el uso de diccionarios de sinónimos y co-ocurrencias (Salton 1968).

Posteriormente, Karen Spärck-Jones (1972) introdujo la idea de que un término no sólo es relevante si aparece frecuentemente en un texto sino que es más valioso cuanto más raro, esto es, cuanto menor es el número de documentos de la colección en que aparece. Esto es lo que se conoce como ***idf*** (*inverse document frequency*[2]) que, al combinarse con la frecuencia de aparición de los términos (Luhn 1957), ha dado lugar a uno de las formas de ponderación de términos más conocidas y utilizadas, *tf*idf*, según la cual la relevancia de un término es directamente proporcional a la frecuencia de aparición en un documento e inversamente proporcional al número de documentos en que aparece. Hay que decir, sin embargo, que Maron y Kuhns (1960, p. 230) ya apuntaron en esa dirección aunque no incidieron en la posibilidad de obtener esos valores de forma automática[3]. Poco después, Stephen E. Robertson y Spärck-Jones (1976) desarrollarían un modelo probabilista para ponderar la relevancia de los términos de una consulta.

Así pues, a finales de los 70, tras dos décadas de investigación, el campo contaba con unas bases teóricas sólidas que ofrecían diversas técnicas para el desarrollo de sistemas de recuperación de información con un rendimiento adecuado. Hay que decir que hasta entonces todos estos sistemas se habían diseñado y evaluado con colecciones de documentos de tamaño conocido[4]; sin embargo, eso iba a cambiar.

2 Internet y la sobrecarga de información

En 1969 comenzó a operar la red *ARPANET* que evolucionaría en los años 80 hasta convertirse en lo que hoy conocemos como Internet. Esta última, al acomodar otras redes de intercambio de información (como *USENET*[5]) y ofrecer soporte para la Web (que, finalmente, ha dado acceso a la práctica totalidad de servicios integrados en Internet) ha

---

[1] Un algoritmo de *stemming* o *stemmer* determina la raíz morfológica de una palabra colapsando múltiples formas de la raíz en un único término (`research`, `researcher`, `researching` y `researchers` colapsan en `research` empleando un *stemmer* para inglés). Un *stemmer* para castellano, por ejemplo, transformaría `andanzas` en `and`, `habitaciones` en `habit` o `juguéis` en `jug`.

[2] El término *inverse document frequency* se ha traducido como "frecuencia inversa del documento", "frecuencia documental inversa", "frecuencia inversa de documentos" o "frecuencia inversa en el documento". Sin embargo estas traducciones son incorrectas y, peor aún, confusas. La medida *idf* trata de ponderar el valor informativo de un término basándose en su frecuencia de aparición en distintos documentos de una colección, a mayor frecuencia de uso menor valor y viceversa. Por ejemplo, `una` aparece en 5 millones de páginas web españolas frente a las 3 que mencionan `folksonomía`; está claro, que el primer término es mucho menos informativo que el segundo. Dicho de otro modo, el valor de un término es inversamente proporcional al número de documentos en que aparece.

[3] *"Un término índice aplicado a cada documento de la biblioteca no tendrá significatividad, mientras que uno aplicado a sólo un documento será altamente significativo. Así pues, las medidas de la significatividad están relacionadas con un "indicador de extensión" para cada término, es decir, con el número de documentos etiquetados con el término –cuanto más pequeño sea este número, mayor será la significatividad del término índice.)"* (Maron y Kuhns 1960, p. 230)

[4] Para el proyecto *Cranfield II* se elaboró una colección con 1.400 documentos (Cleverdon y Keen 1966), la colección *NPL* contenía 11.429 (Spärck-Jones y Webster 1979) y la *CACM* 3.204 (Fox 1983).

[5] *USENET* permite a los usuarios publicar, leer y comentar mensajes de texto (artículos) sobre distintos temas organizados en grupos jerárquicos.

contribuido en enorme medida, junto con el tremendo abaratamiento del soporte en disco magnético[1], a la explosión de información textual que se vive hoy en día.

Maron y Kuhns (1960, p. 217) afirmaban que los datos documentales estaban siendo generados a un ritmo alarmante. Cuatro décadas después es difícil encontrar un superlativo para "alarmante" que dé idea de la tasa de producción de información que ha alcanzado la humanidad, tan sólo algunas cifras:

- Desde 1981 se han generado más de 845 millones de mensajes en *USENET*[2], lo cual supone una media de más de 100.000 mensajes nuevos al día.
- La Web superficial consta, al menos, de 4.000 millones de documentos[3], la Web oculta (Florescu, Levy y Mendelzon 1998), es decir, aquellas páginas accesibles sólo tras rellenar algún tipo de formulario, tendría según Isidro F. Aguillo (2002) entre 2 y 50 veces el tamaño de la primera y según Michael K. Bergman (2001) hasta 500.
- *Reuters* produce alrededor de 11.000 artículos de prensa diarios[4] (aproximadamente 2,5 millones de palabras) que ofrece de manera *online* mediante documentos *NewsML*[5].
- *Springer Verlag* editó 356 volúmenes de su serie *Lecture Notes in Computer Science* en 2003[6] (alrededor de 90 millones de palabras). Todos los artículos están disponibles como archivos *PDF*[7]).

La lectura de estos datos parece evocar una mezcla de imágenes: un ejército de tipógrafos escribiendo, no los volúmenes de la Biblioteca del Museo Británico, sino de la Biblioteca de Babel. Como se verá, la realidad no está muy lejos de la ficción.

Actualmente cualquier usuario puede publicar en *USENET* y, a poco que lo desee, en la Web lo cual supone alrededor de 400 millones[8] de potenciales "efectivos". Según un estudio realizado en 2003 por *Pew/Internet* en EE.UU. entre usuarios adultos (18 años o más) un 44% ha contribuido de una forma u otra a incrementar los contenidos disponibles[9], el 10% ha publicado en grupos de noticias (*USENET*), el 13% mantiene sitios web y el 2% bitácoras[10] (Lenhart, Horrigan y Fallows 2004).

---

[1] Una unidad de disco *IBM 1301* tenía una capacidad aproximada de 27MB y costaba, en 1961, 115.500 dólares (*IBM* 1994-2004). Un disco de 40GB de la misma marca costaba en 2004 170 dólares. Teniendo en cuenta la inflación (Sahr 2004), el precio del primer sistema equivaldría a 725.000 dólares actuales por lo que el coste por megabyte ha descendido desde 26.850 dólares a 0,4 centavos en 43 años.

[2] Fuente: *Google* (`http://www.google.com`)

[3] Fuente: *Google* (`http://www.google.com`)

[4] Fuente: *Reuters* (`http://www.reuters.com`)

[5] *NewsML* (*News Markup Language*, Lenguaje de Etiquetado de Noticias) es un vocabulario XML para "empaquetar" contenidos periodísticos así como para añadirles metainformación.

[6] Fuente: *Springer Verlag* (`http://www.springeronline.com`)

[7] *Portable Document Format* (Formato de Documento Transportable).

[8] En 2001 había alrededor de 119 millones de usuarios de Internet en la Unión Europea (David 2003), 143 en EE.UU. (*USDOC* 2002), 47 en Japón (*MPHPT* 2001) y, en 2003, 68 millones en China (*CNNIC* 2003).

[9] Compartiendo archivos, anotando comentarios en sitios web y bitácoras, respondiendo artículos en *USENET*, etc.

[10] Una bitácora, *weblog* o su contracción *blog* es un sitio web que contiene, en orden cronológicamente inverso, artículos publicados por una persona (en ocasiones por grupos) sobre los más diversos temas empleando un sistema de gestión de contenidos. Las bitácoras tienen dos características esenciales, suelen

De acuerdo con (*USDOC* 2002) de los 143 millones de usuarios en EE.UU. 95 millones son adultos. Combinando este dato con el informe anterior tendríamos que, sólo en dicho país, 10 millones de usuarios publican en *USENET*; 12 millones mantienen sitios web y casi 2 millones bitácoras. Si las mismas ratios demográficas y actitudes de uso fuesen extensibles a la "población total" de Internet tendríamos 27 millones de usuarios que habitualmente publican en *USENET*, 35 millones de *webmasters* y 5 millones y medio de *bloggers*.

Naturalmente, los datos no son directamente extrapolables aunque es razonable suponer que Europa y Japón exhiben comportamientos similares. Por ello, la cifra de 5 millones de creadores habituales de contenido textual parece bastante razonable como cota inferior y la imagen de un formidable ejército tecleando creíble.

Por otro lado, la Web[1] se asemeja bastante a la Biblioteca que describió Borges. Su tamaño real es desconocido. No todos los contenidos disponibles son realmente interesantes[2] y, lo que es más, a pesar de que muchas personas, en particular estudiantes (Grahan y Metaxas 2003), lo den por hecho tampoco son necesariamente veraces.

Sin embargo, ni el tamaño de la Web ni la veracidad de sus contenidos son problemas que se vayan a abordar en este trabajo sino uno más específico, a saber, la sobrecarga de información que sufren todos los usuarios que utilizan la Web como fuente de información. No obstante, para ahondar en la naturaleza del problema es necesario analizar por qué y cómo surgió la Web, cuál es el motivo por el que crece a un ritmo tan elevado y qué soluciones se han propuesto para localizar información en el aluvión disponible.

3 La Web como sistema de recuperación de información

Tim Berners-Lee (1989) inició el desarrollo de la Web en el *CERN*[3] como un medio para evitar la pérdida de información, inevitable en una organización de gran tamaño, y facilitar el acceso a la información disponible (bases de datos, directorios telefónicos, etc.) Dos características de la propuesta original permitieron transformar el proyecto original en la Web actual: su naturaleza distribuida (los documentos pueden residir en máquinas distintas) y la posibilidad de establecer vínculos (enlaces) entre documentos. Por otro lado, Berners-Lee insistía en la necesidad de construir un sistema que animase a los usuarios a incorporar nueva información, haciéndolo así aun más útil y atractivo, de tal forma que el conjunto de documentos creciese de forma continua.

Berners-Lee hace algunas reflexiones muy interesantes sobre posibles problemas y métodos para recuperar información en un sistema como el que proponía. Alerta, por ejemplo, sobre los inconvenientes de utilizar palabras clave para localizar documentos y sugiere la posibilidad de establecer enlaces, no sólo con documentos, sino también con

---

ser muy personales y los lectores pueden anotar los artículos con sus propios comentarios. La persona que mantiene una bitácora es un *blogger* y la totalidad de *blogs* es conocida como blogosfera.

[1] Entendiendo la Web como el conjunto formado por la Web superficial, la oculta y todos los servicios disponibles y/o accesibles vía web ahora y en el futuro (bitácoras, *USENET*, periódicos digitales, publicaciones científicas en formato electrónico, enciclopedias, foros, etc.)

[2] Como diría Borges *"por una línea razonable o una recta noticia hay leguas de insensatas cacofonías, de fárragos verbales y de incoherencias"*.

[3] *Centre Européen pour la Recherche Nucléaire*, Centro Europeo para la Investigación Nuclear.

conceptos facilitando la existencia de enlaces "indirectos" entre documentos de temática similar.

Por tanto, la propuesta original planteaba construir la Web sobre una base semántica más o menos sólida, partiendo de nodos conceptuales enlazados desde los distintos documentos. Por otra parte, para explotar las ventajas de los enlaces indirectos antes mencionados, los enlaces entre documentos y conceptos deberían ser bidireccionales. No obstante, Berners-Lee (1989) no hace ninguna mención explícita sobre enlaces bidireccionales, tan sólo en (Berners-Lee 1990) se plantea su utilidad aunque también señala que un programa que recorriese una Web monodireccional podría obtenerlos simplemente "invirtiendo" los enlaces que hubiese encontrado. Finalmente, las primeras versiones del lenguaje *HTML*[1] (*HyperText Mark-up Language* 1992) establecieron de manera permanente la unidireccionalidad de los enlaces, esto es, si un documento *A* enlaza a un nodo *B* es posible "llegar" a *B* partiendo de *A* e imposible[2] hacer el recorrido contrario comenzando en *B*.

Además, en ninguna de las versiones desarrolladas se incluyó nada similar a los "nodos conceptuales". Las primeras versiones del lenguaje *HTML* (*HyperText Mark-up Language* 1992) permitían dar título a los documentos, formatear texto (párrafos, listas, cabeceras, etc.), crear enlaces a otros recursos y, de un modo rudimentario mediante la etiqueta `ISINDEX`, crear interfaces con programas auxiliares para realizar búsquedas de documentos en cada servidor (mediante palabras clave). Como consecuencia, la Web se convirtió en un artefacto diseñado para crecer de un modo cada vez más acelerado sin proporcionar mecanismos adecuados para localizar información[3].

4 Los primeros directorios y motores de búsqueda[4]

El primer servidor web (`info.cern.ch`, antes `nxoc01.cern.ch`) entró en funcionamiento en 1990 y para finales de 1992 existían alrededor de veinte (Berners-Lee 1992c). En aquel momento resultaba relativamente sencillo mantener de manera manual un directorio de sitios web y, de hecho, organizaciones como el *CERN* o el *NCSA*[5]

---

[1] *HyperText Mark-up Language* (Lenguaje de Etiquetado Hipertextual) es un lenguaje de etiquetas basado en *SGML* que permite construir páginas web. El estándar para este lenguaje es mantenido por el *Consorcio W3* (`http://www.w3.org`).

[2] Es cierto que existen herramientas como *Google* (`http://www.google.com`) que permiten encontrar documentos que enlazan con uno específico (por ejemplo, la consulta `link:http://www.w3.org` proporciona como resultado documentos que enlazan con el sitio del *Consorcio W3*). Sin embargo, esto es posible tan sólo después de haber explorado una porción de la Web construyendo un grafo que la represente y permita un análisis posterior como sugería Berners-Lee (1990), no se trata de una característica intrínseca de la Web.

[3] Tim Berners-Lee (1992a) sugiere utilizar documentos índice (aquellos que emplean la etiqueta `ISINDEX`) que, a su vez, apuntasen a otros índices más específicos y en (Berners-Lee 1992b) plantea utilizar la estructura formada por los enlaces para encontrar información relevante. Tan sólo son unos breves apuntes y las ideas están esbozadas de un modo rudimentario pero señalan las dos líneas que posteriormente se seguirían para desarrollar mecanismos de búsqueda en la Web: índices (por ejemplo, *Yahoo!*) y buscadores basados en robots (por ejemplo, *Google*). A pesar de todo, los métodos para localizar información en la Web son externos a la misma.

[4] Un motor de búsqueda, o simplemente buscador, es un artefacto *software* que explora la Web almacenando en una base de datos parte o todo el texto de los documentos que analiza. Al ir procesando documentos se crea un índice que emplea las palabras que aparecen en cada página web. Cuando un buscador recibe una consulta toma las palabras utilizadas por el usuario y obtiene los documentos indexados por las mismas.

[5] *National Center for Supercomputing Applications*, Centro Nacional para Aplicaciones de Supercomputación.

gestionaban índices a los que iban añadiendo las notificaciones de nuevos sitios que recibían por correo electrónico.

Sin embargo, según un estudio realizado por Matthew Gray (1995), a finales de 1993 había 623 servidores web y su número se duplicaba cada 3 meses, existiendo en diciembre de 1994 más de 10.000. Esto hacía muy difícil, aunque no imposible, mantener manualmente un índice de sitios web, dejando a un lado el hecho de que muchos administradores no notificaban su existencia a los directorios existentes (Steinberg 1996, p. 2). Así, la carencia de un sistema adecuado para poder localizar los distintos servidores y documentos en la incipiente Web era ya un problema[1] y comenzaron a desarrollarse distintos sistemas en busca de soluciones.

Tan sólo en *WWW94* (*First International Conference on the World-Wide Web*, Primer Congreso Internacional sobre la Web) se presentaron nueve trabajos relativos al indexado automático de documentos y a la búsqueda de información. Entre ellos cabe destacar los sistemas creados por Martijn Koster (Koster 1994) y Oliver A. McBryan (McBryan 1994), *ALIWEB* y *WWW Worm*, respectivamente. Ambos desarrollaron programas para explorar la Web de manera automática[2], saltando de enlace en enlace y almacenando información sobre las páginas visitadas en una base de datos para su posterior consulta por parte de los usuarios.

ALIWEB (Koster 1994) comenzaba su exploración a partir de sitios web registrados manualmente por sus administradores almacenando una información relativamente escasa para cada documento indexado (título, descripción y algunas palabras clave) lo que limitaba las posibilidades de los usuarios al realizar sus consultas[3].

En el caso de *WWW Worm* (McBryan 1994) no queda muy claro cómo se construía la base inicial de sitios web para realizar el indexado, la información almacenada era aún más parca (título del documento y textos utilizados en los enlaces que apuntan al mismo) y se consultaba (internamente) mediante la orden *UNIX* `egrep`[4].

Otros sistemas destacables, similares a los anteriores y desarrollados en la misma época fueron *Jumpstation*, *Wanderer*, *WebCrawler* y *Lycos*.

Jumpstation, implementado por Jonathon Fletcher (1994), fue uno de los primeros motores de búsqueda. Entró en funcionamiento en diciembre de 1993 y lo hizo de manera errática hasta quedar desatendido en abril de 1994.

---

[1] A lo largo de 1993 tuvo lugar una interesante discusión en la lista de correo *WWW Talk* sobre distintas formas de enfocar la localización de recursos en la Web así como los retos que plantearía: recorrer todos los enlaces almacenando la información encontrada (Perry 1993), la necesidad de grandes recursos temporales y de almacenamiento así como la conveniencia de no visitar todos los enlaces posibles (Johnson 1993) o los problemas que surgirían con enlaces generados automáticamente (Putz 1993). Por otro lado, Thomas R. Bruce (1993) planteaba un enfoque no automático mediante el cual los distintos sitios web deberían "registrarse" para ser revisados por "editores" que los categorizarían, pudiendo dividirse las categorías en subgrupos que serían asignados a nuevos editores; el mismo planteamiento que, algunos años después, sigue *ODP* (*Open Directory Project*, Proyecto de Directorio Abierto, `http://www.dmoz.org`)

[2] *Spiders*, "arañas", también denominados robots.

[3] La probabilidad de coincidencia entre dos individuos (por ejemplo, entre el autor de un documento y un potencial lector) en el uso de la misma palabra para identificar un concepto está entre el 10 y el 20% (Furnas *et al.* 1987).

[4] `egrep` permite realizar búsquedas en un fichero de texto empleando expresiones regulares, como resultado muestra todas las líneas que contienen el patrón recibido como parámetro.

Wanderer fue inicialmente desarrollado para descubrir nuevos sitios web, posteriormente se utilizó para medir la expansión de la Web entre junio de 1993 y junio de 1995 (Gray 1995) y, finalmente, para construir el buscador *Wandex* (*Wanderer Index*).

WebCrawler (Pinkerton 1994) supuso una mejora respecto a *ALIWEB* o *WWW Worm* puesto que indexaba todo el texto de las páginas que exploraba. Esta estrategia permitía ofrecer más documentos para las consultas de los usuarios pero, al aumentar el número de páginas indexadas, reducía de manera drástica la **precisión**[1] de las respuestas.

Lycos (Mauldin y Leavitt 1994) constituyó una iniciativa intermedia entre *ALIWEB* y *WebCrawler* puesto que no indexaba el texto completo de los documentos ni únicamente su título y descripción. En su lugar generaba una versión "ligera" constituida por el título, las veinte primeras líneas y las cien palabras más relevantes[2].

Este tipo de sistemas automáticos podían enfrentarse al enorme crecimiento de la Web en mejores condiciones que los índices construidos de forma manual. Sin embargo, estos últimos ofrecían otras ventajas (organización en categorías jerárquicas, posible revisión por parte de "editores" especializados, referencias cruzadas, etc.) que también eran valoradas por los usuarios (Bruce 1993). Por ejemplo, *Galaxy, Yahoo!* y *ODP* se construyeron siguiendo esta línea.

Galaxy (Speyer y Allen 1994) empezó a dar servicio a comienzos de 1994. Según el anuncio original se habían empleado métodos semiautomáticos para construir la base de datos original (que incluía no sólo páginas web sino también servidores *Gopher*[3] y *WAIS*[4]). *Galaxy*, como sería común en todos los directorios posteriores, permitía que los administradores de sitios web notificasen la dirección de su sitio para su inclusión en una categoría previamente seleccionada entre las disponibles en la jerarquía, posteriormente, un editor revisaba el sitio web y decidía acerca de su inclusión en el directorio.

El sitio web precursor de *Yahoo!* (Filo y Yang 1994), *Jerry's Guide to the World Wide Web* – Guía de Jerry a la Web, fue creado a comienzos de 1994 como un proyecto personal y se transformó en un directorio comercial en 1995. Al igual que *Galaxy*, dispone de categorías predefinidas en las que los administradores de sitios web pueden solicitar la inclusión que, también, es revisada por empleados de la empresa.

ODP[5] (*Open Directory Project*, Proyecto de Directorio Abierto), antes *DMoz* (*Directory Mozilla*, Directorio Mozilla), antes *NewHoo* y, aún antes, *GnuHoo*[6], fue fundado en 1998. Su

---

[1] La precisión es la proporción entre el número de documentos relevantes retornados por un sistema para una consulta y el total de documentos retornados.

[2] Aplicando *tf*idf* como método de ponderación (véase página 4).

[3] *Gopher* es un protocolo para la búsqueda y recuperación de información textual en Internet, fue desarrollado en la Universidad de Minnesota (EE.UU.) y distribuido públicamente a partir de 1991. El objetivo de *Gopher* era muy similar al de la Web aunque resultaba menos flexible que HTML. Todavía existen algunos servidores *Gopher* activos (menos de 300 en Agosto de 2004, fuente: *Floodgap* <http://gopher.floodgap.com>) pero puede afirmarse que ha sido totalmente reemplazado por la Web.

[4] *WAIS* (*Wide Area Information Servers*, Servidores de Información a Nivel Global) era un sistema para búsqueda remota de texto en bases de datos distribuidas basado en el estándar *ANSI Z39.50*. Hasta donde sabe el autor no existe ningún servidor *WAIS* operativo en la actualidad.

[5] http://www.dmoz.org

[6] El cambio de *GnuHoo* a *NewHoo* estuvo motivado, aparentemente, por un artículo (Miller 1998) publicado en *Slashdot* que criticaba el uso del acrónimo *GNU* (vinculado a proyectos de *software* no

estructura y funcionamiento es similar a la de *Galaxy* o *Yahoo!* con la diferencia de que los editores no forman parte de la plantilla de la empresa sino que realizan su labor de manera desinteresada.

Así pues, en torno a 1998 ya existía toda una serie de recursos para la búsqueda de información en la Web que, sin embargo, seguían siendo insuficientes. Por un lado, no parecía que los directorios como *Yahoo!*, al utilizar editores humanos, pudiesen clasificar todos los sitios web existentes (mucho menos todas las páginas) al mismo ritmo que aparecían (Steinberg 1996). Por otro lado, aunque algunos expertos[1] afirmaban que era posible indexar la Web al mismo ritmo que crecía, otros, como Steve Lawrence y C. Lee Giles (1998), discrepaban. Éstos, tras analizar la "cobertura" de distintos buscadores[2] encontraron que ninguno cubría, individualmente, más de un tercio de la Web indexable[3] conocida y que la combinación de varios buscadores (en su caso seis) podía cubrir más del triple de páginas que un único sistema.

Lawrence y Giles (1998) concluyen que el uso de metabuscadores era una solución para localizar información en la Web puesto que garantizaba la cobertura del mayor número posible de páginas. Uno de los primeros metabuscadores, de hecho anterior a su estudio, fue *MetaCrawler* (Selberg y Etzioni 1995). Selberg y Etzioni señalan dos deficiencias fundamentales en los buscadores de aquel momento: (1) la porción de la Web sobre la que trabaja cada buscador es distinta del resto obligando a los usuarios a repetir la consulta en distintos buscadores y (2) gran parte de los resultados son irrelevantes o enlaces "muertos". *MetaCrawler* pretendía dar solución al primer problema ofreciendo una interfaz única para los distintos buscadores (esto es, distintas bases de datos) y el segundo filtrando los resultados recibidos.

Hay que señalar, sin embargo, que ni los argumentos que se presentan para señalar el primer problema ni los criterios que emplean para "evaluar" la relevancia de los documentos son totalmente acertados (Lawrence y Giles 1998, p. 98). Esto en absoluto invalida la propuesta de Selberg y Etzioni; de hecho, los metabuscadores siguen siendo comunes[4]. Sin embargo, no son la solución al problema de la relevancia ni tampoco al de la cobertura puesto que, en rigor, bastaría un solo buscador común que indexara un número suficiente de páginas web.

5 Motores de búsqueda modernos

En 1998 los buscadores existentes no cubrían, individualmente, toda la Web ofreciendo los metabuscadores una solución simple e inmediata para ese problema (Lawrence y Giles 1998). Sin embargo, el problema también podría solventarse si los

---

privativo) por parte de una empresa privada para denominar una alternativa a *Yahoo!* construida mediante el uso de voluntarios en lugar de personal propio.

[1] Eric A. Brewer, fundador y responsable técnico de *Inktomi* (actualmente parte de *Yahoo!*) afirmó en una entrevista (Steinberg 1996, p. 5) que era posible indexar la totalidad de la Web, al menos, hasta 2000.

[2] Los buscadores que analizaron fueron *AltaVista*, *Excite* (que había adquirido *WebCrawler*), *HotBot*, *Infoseek*, *Lycos* y *Northern Light*. Ninguno era un directorio, todos utilizaban robots.

[3] Aquella parte de la Web accesible para un robot, es decir, páginas web accesibles sin formularios, contraseñas o que no les están "vedadas" (existen una serie de estándares para evitar que los robots de búsqueda no accedan a ciertas partes de un sitio web).

[4] En Marzo de 2004 existían alrededor de 30 metabuscadores activos. Fuente: *SearchEngineWatch* <http://searchenginewatch.com>

buscadores cubriesen porciones de la Web mayores (idealmente su totalidad), algo teóricamente factible.

Por otro lado, no es de extrañar que búsquedas realizadas sobre una base documental con, al menos, 320 millones de *ítems* (Lawrence y Giles 1998) proporcionasen resultados de una relevancia mediocre al realizar búsquedas por palabras clave con esquemas de ponderación muy simples. Sin embargo, gracias a la naturaleza del hipertexto, podía tratar de encontrarse una solución empleando métodos análogos a los aplicados a un tema antiguo: la cita de trabajos científicos.

Una característica de los textos científicos es la referencia a trabajos de terceros. El estudio de los patrones subyacentes a las referencias en revistas científicas ha llegado a convertirse en una rama del conocimiento con, al menos, 75 años de antigüedad (Lotka 1926), (Gross y Gross 1927), (Brodman 1944) o (Fussler 1949) y hace algo más de treinta que se desarrolló el concepto de "índice de impacto" (Garfield 1972) para determinar el "prestigio" de las distintas publicaciones.

La idea tras dicho índice es muy simple: partiendo de una base de datos que almacene para cada trabajo científico su título, la revista en que fue publicado y la lista de obras que cita es posible obtener el número de referencias hechas a un trabajo, autor o revista. Obviamente, los trabajos más antiguos, los autores más prolíficos o las revistas con más artículos o números por año serán, probablemente, más citados a pesar de que otros trabajos, autores y publicaciones pueden ser tanto o más relevantes que los primeros. Así pues, es necesario algún tipo de "normalización" del número total de citas a fin de obtener una medida de índice de impacto "justa".

Sin embargo, no es necesario entrar en mayores detalles para encontrar paralelismos entre el problema de la relevancia de los trabajos científicos y la relevancia de los documentos en la Web. En el primer caso por medio de las referencias y en el segundo mediante los enlaces se establecen vínculos entre documentos que indican, implícitamente, que el autor que establece dicho vínculo considera al documento enlazado tanto o más relevante que el suyo propio[1] o una fuente autorizada sobre un tema en particular. Por otro lado, el contexto en que se hace la cita (o el texto que rodea y se utiliza en el enlace) aportan información muy valiosa sobre el contenido del documento referenciado. Ideas similares a estas podían ser aplicadas inmediatamente a las bases de datos construidas por los robots al explorar la Web aprovechando las características del hipertexto.

Jon Kleinberg (1998) sentó las bases sobre las que se apoyan los modernos buscadores al plantearse la viabilidad de un método algorítmico para estimar la relevancia de un documento, algo que según él era una característica subjetiva. Para ello definió los conceptos de **"autoridad"** y ***hub*** (concentrador). Una autoridad es un documento al que enlazan muchos otros puesto que, según Kleinberg, cada enlace recibido es un "voto"

[1] En algunas ocasiones se referencia un trabajo para criticarlo duramente o se establece un enlace a un sitio web por motivos maliciosos (véase más adelante). Puesto que el uso pretendido, y habitual, de referencias y enlaces tiene connotaciones positivas el abuso tiene efectos perturbadores. En el primer caso (referencias científicas) un trabajo polémico puede ser muy citado y, consecuentemente, valorado como muy relevante. En el segundo se puede abusar de los mecanismos de un motor de búsqueda para construir complicadas "bromas". Por ejemplo, al formular la consulta `ladrones` en *Google* se obtiene como primer resultado el sitio web de la *Sociedad General de Autores y Editores* [17 agosto 2004]. Esto se conoce como *Google Bomb* – Bomba *Google* y requiere un esfuerzo coordinado de varios sitios web distintos para lograr su objetivo. Este tipo de hechos, a pesar de encerrar cierta malicia, son difícilmente clasificables como "ataques".

emitido por el individuo que estableció dicho enlace. Analizando el texto empleado en los enlaces puede determinarse el contexto en el cual el documento enlazado es una autoridad. Por su parte, un concentrador será un documento que contiene enlaces a muchas autoridades y es, por tanto, un recurso valioso para localizar información relevante en la Web.

Estos conceptos fueron probados por Chakrabarti *et al.* (1998a) y (1998b) mediante varios prototipos que tenían como objetivo localizar únicamente los documentos más relevantes para cada consulta, esto es, las autoridades. Para evaluar el rendimiento de estas técnicas se realizaron una serie de consultas genéricas empleando dichos prototipos, *Yahoo!* (un directorio) y *Altavista* (un buscador basado en robots) obteniendo, en cada caso, los diez documentos más relevantes. Posteriormente, un grupo de usuarios evaluó de manera "ciega" cada documento y valoró su relevancia en relación con la consulta planteada. La relevancia media de los resultados proporcionados empleando la técnica de Kleinberg superaba el 50% frente al 40% de *Yahoo!* y el 20% de *Altavista* abriendo la posibilidad de construir automáticamente taxonomías de documentos similares a las construidas por expertos humanos.

Existen, sin embargo, tres escenarios (Bharat y Henzinger 1998) en los que la técnica de Kleinberg puede ser objeto de abuso o simplemente falla por basarse en suposiciones no totalmente correctas. Se trata de relaciones entre servidores "mutuamente fortalecedoras", enlaces generados automáticamente y documentos irrelevantes enlazados desde autoridades o concentradores.

Relaciones entre servidores "mutuamente fortalecedoras". El algoritmo de Kleinberg cuenta cada enlace como un voto diferente; de este modo, si varios documentos alojados en un único servidor apuntan a un único documento externo éste recibe muchos "votos" que Bharat y Henzinger consideran "fraudulentos". Para solucionarlo plantean la necesidad de reducir el peso otorgado a los enlaces que parten desde un único servidor a un único documento. No obstante, puesto que un servidor puede alojar múltiples sitios web (páginas personales por ejemplo) su solución es simplista y devalúa "votos" independientes por cuestiones meramente topológicas. Brian D. Davison (2000a) realiza un estudio mucho más riguroso sobre el difícil problema de los enlaces "nepotistas".

Enlaces generados automáticamente. Según Kleinberg un enlace es un "voto" emitido por un individuo a favor de la relevancia de un documento. Sin embargo, existen enlaces que no son creados por seres humanos sino generados automáticamente con lo cual ya no son "votos" válidos. Por ejemplo, al crear una página personal o una bitácora en algún servidor gratuito todos los documentos tendrán enlaces a la página principal del servicio o de los patrocinadores. Tales enlaces no pueden diferenciarse de los creados por una persona y se valorarán de igual modo, afectando de un modo difícil de determinar a los resultados. Bharat y Henzinger no ofrecen solución a este problema.

Documentos irrelevantes enlazados desde autoridades o concentradores. Se supone que si una autoridad o un concentrador enlazan un documento éste debe ser necesariamente una autoridad sobre el tema tratado en los documentos de partida. Sin embargo, esta suposición no siempre es cierta; por ejemplo, las páginas personales de los autores de documentos muy referenciados no tienen por qué ser necesariamente relevantes. Bharat y Henzinger proponen analizar el contenido de las páginas enlazadas para comprobar si realmente tienen relación con el tema tratado en el documento del que parte el enlace. Para mejorar el rendimiento de su analizador no emplean palabras clave sino raíces obtenidas mediante el algoritmo de *stemming* de Porter (1980). Hay que decir que la idea de

analizar los contenidos es atractiva pero lo cierto es que este planteamiento es poco escalable en un entorno multilingüe como la Web (habría que desarrollar un *stemmer* para cada idioma).

En 1998 comenzó a operar el buscador, tal vez, más popular de la actualidad: *Google* (Brin y Page 1998). Éste, al igual que los motores de búsqueda "tradicionales", emplea robots para explorar la Web en búsqueda de documentos pero, al contrario que estos, utiliza una técnica mucho más sofisticada para organizar los resultados de las consultas de los usuarios: el algoritmo ***PageRank*** (Page *et al.* 1998), similar en ciertos aspectos al propuesto por Kleinberg. Al igual que Kleinberg, *PageRank* se basa en el uso de autoridades; sin embargo, no todos los enlaces son valorados del mismo modo sino en función de un valor numérico otorgado a cada documento, denominado también *PageRank*. Dicho valor indica el "prestigio" o la relevancia del documento y se propaga de unos documentos a otros: el *PageRank* de una página se divide por el número de enlaces de salida y se "transfiere" a los documentos enlazados. Así, documentos que reciben muchos enlaces aunque de poco valor serán muy relevantes y documentos que reciben pocos enlaces pero desde páginas con *PageRank* elevado serán igualmente importantes.

Además del valor *PageRank*, *Google* utiliza otros factores para ordenar los resultados de una consulta, por ejemplo, el texto de los enlaces que reciben los documentos, la posición de las palabras clave dentro del documento, etc. De este modo, los primeros documentos se corresponden, aproximadamente, con las autoridades que se obtendrían aplicando el algoritmo de Kleinberg pero sin eliminar la posibilidad de consultar otros documentos con menor *PageRank*. De este modo *Google* recupera muchos documentos para cada consulta (en ocasiones cientos o miles) pero ofreciendo siempre los documentos más relevantes[1] entre los primeros resultados.

Este sistema parece adecuado para la mayor parte de usuarios. Según Jansen *et al.* (1998) y Silverstein *et al.* (1998) los usuarios de motores de búsqueda resuelven una necesidad de información con menos de dos consultas (en un 67% de los casos de acuerdo con el primer estudio y en un 78% según el segundo), no suelen pasar de la primera página de resultados (en un 58% según los primeros y en un 85% según los segundos) y, de acuerdo con Jansen y Spink (2003), un 66% de los usuarios examinan entre los resultados menos de 5 documentos y un 30% un único documento. Jansen y Spink argumentan que esto se debe a tres razones: (1) las necesidades de información de la mayoría de internautas no son complejas, (2) los primeros documentos retornados son realmente "autoridades" para la consulta formulada y, (3) en promedio, alrededor del 50% de los documentos retornados son relevantes para una consulta específica desde la perspectiva del usuario (Jansen y Spink 2003, p. 68).

Esto suscita varias cuestiones. ¿Por qué retornar entonces miles de documentos para cada consulta? ¿No serían suficientes los *n* más relevantes a la manera de Kleinberg? ¿En cuántos casos no se alcanza una precisión del 50% en la primera página de resultados? ¿Existen usuarios que no encuentran lo que buscan entre los documentos con mayor puntuación pero que podrían hacerlo en alguno de los centenares apenas puntuados?

Tratando de dar respuestas a estas preguntas el autor de este trabajo realizó un sencillo experimento que se pasa a describir. En primer lugar, se recopilaron tres conjuntos de consultas. El primero estaba formado por las diez consultas más frecuentes realizadas en *Google* por usuarios españoles durante julio de 2004 (véase Tabla 1). El segundo por las 50

[1] Los más relevantes según el criterio de *Google* que no siempre coincidirá con el criterio del usuario.

consultas más frecuentes, según *Wordtracker*[1], realizadas el 17 de agosto de 2004 (véase Tabla 2). Y el tercero por 50 consultas extraídas en tiempo real de *Dogpile – Searchspy*[2], también el 17 de agosto (véase Tabla 3).

Posteriormente, cada consulta era enviada a *Google* y se obtenían los siguientes datos:

- Número de resultados totales.
- Aparición de un sitio web "oficial" como primer resultado (podría indicar que se trata de una consulta navegacional).
- En caso de no existir sitio web "oficial", número de temas distintos tratados por los documentos resultantes (podría dar pistas sobre la ambigüedad de la consulta).
- Porcentaje de documentos relevantes[3] para la consulta entre los diez primeros resultados. En caso de que apareciese un sitio web "oficial" como primer resultado se asignaba una relevancia del 100% al considerar que la consulta era "navegacional", esto es, tan sólo pretendía alcanzar un sitio web cuya existencia era conocida por el usuario.

Consulta	Sitio web oficial	Temas distintos	Resultados	Relevancia
shrek 2	✓	-	2.260.000	100%
paginas amarillas	✓	-	383.000	100%
renfe	✓	-	320.000	100%
sport	✓	-	171.000.000	100%
iberia	✓	-	952.000	100%
el corte ingles	✓	-	387.000	100%
harry potter	✓	-	6.280.000	100%
hola	✓	-	1.870.000	100%
chistes		**1**	**2.460.000**	**100%**
postales		1	4.350.000	100%

Tabla 1. Diez consultas más frecuentes realizadas por internautas españoles en Julio de 2004 (Fuente: *Google*).

A partir de los datos recogidos en Tabla 1 y Tabla 2 puede concluirse que gran parte de las consultas más frecuentes se corresponden con uno de los tres tipos siguientes[4]: (1) entretenimiento (por ejemplo, `shrek 2`, `outback jack` o `chistes`), (2) servicios disponibles

[1] `http://www.wordtracker.com`

[2] `http://www.dogpile.com/info.dogpl/searchspy`

[3] Cada una de las páginas obtenidas como resultado fue visitada por el autor para determinar si era o no relevante (según criterios humanos) para la consulta formulada.

[4] Las categorías propuestas por el autor de este trabajo no se corresponden con las de la taxonomía habitualmente utilizada (Broder 2002). De hecho, las consultas de "entretenimiento" y "servicios" pueden ser navegacionales (por ejemplo, `shrek 2` o `hotmail`), transaccionales (por ejemplo, `juegos` o `weather`) e incluso informativas (por ejemplo, `chistes` o `jokes`) mientras que las de "celebridades" podrían ser navegacionales (por ejemplo, `avril lavigne` o `pamela anderson`) o informativas (por ejemplo, `paris hilton` o `dan castellaneta`). Dos motivos impulsaron al autor a proponer una taxonomía diferente. Por un lado, sólo se querían clasificar las consultas más frecuentes, no cualquier consulta posible. Por otro, la taxonomía de Broder se basa en las intenciones del usuario que subyacen a la consulta. Por ejemplo, una consulta "navegacional" indica que el usuario conoce la existencia de un sitio web y emplea el buscador para alcanzarlo. Así, desde ese punto de vista la consulta `avril lavigne` es navegacional desde mediados de 2002; antes, al no existir sitio web oficial, la consulta era "informativa", se deseaba encontrar información sobre la artista en una o más páginas. Dada la naturaleza de las consultas más frecuentes parecía más adecuada una taxonomía centrada en el objeto de la consulta que en las intenciones del usuario que la realiza.

vía Web (por ejemplo, `el corte ingles`, `ebay` o `weather`) y (3) celebridades (por ejemplo, `paris hilton` o `ian thorpe`).

En el primer caso las consultas son simples; si se trata de una película, libro o programa de televisión los usuarios emplean el título como consulta (por ejemplo, `harry potter`), si no, se limitan a indicar qué quieren (por ejemplo, `chistes`). Para el segundo tipo o bien se utiliza como consulta el propio nombre del servicio (por ejemplo, `ebay`, `yahoo` o `hotmail`) o palabras clave bien conocidas (`postales`, `jokes`, `weather`, `jobs`, etc.) para describir servicios "genéricos" (el usuario no tiene preferencia por ninguno en particular). Por último, en el tercer caso la consulta se reduce al nombre y/o apellido de la celebridad.

Este tipo de consultas frecuentes parecen obtener resultados satisfactorios de manera regular por tres motivos: existe un sitio web oficial que cubre las necesidades de información, o bien uno o más sitios web que ofrecen servicios análogos (por ejemplo, postales, juegos o información meteorológica) o se trata de información disponible en periódicos digitales[1].

Así, el experimento realizado por el autor parece llegar a las mismas conclusiones que Jansen *et al.* (1998), Silverstein *et al.* (1998) y Silverstein y Spink (2003): las consultas no suelen tener más de dos términos, no es necesario pasar de la primera página de resultados y, en promedio, más del 50% de los primeros resultados son relevantes para la consulta formulada (de hecho, el porcentaje es ligeramente superior al 90% en el caso de las consultas más frecuentes).

De manera análoga se procedió a estudiar las 50 consultas obtenidas en tiempo real que, en teoría, constituyen una muestra razonable de consultas "típicas" pero no frecuentes. Se eliminaron las que obtenían como primer resultado un sitio web "oficial" y aquellas para las cuales fue imposible determinar la relevancia de los resultados[2]. De este modo quedaron 35 consultas que obtuvieron unos resultados con una precisión promedio del 55%.

Ese dato también es coherente con los resultados de Silverstein y Spink (2003). Sin embargo, estos investigadores no indican la dispersión que muestra la relevancia de los resultados en su experimento. Por su parte, el autor ha detectado en esta muestra que el 20% de las consultas obtiene una precisión media de tan sólo el 21% y el 23% de las consultas no obtienen ningún documento relevante entre los 10 primeros.

---

[1] Aparentemente, algunas de las consultas del tercer tipo buscan información sobre "escándalos" recientes en los que estaría envuelta la celebridad en cuestión. Este sería el caso de las consultas `kobe bryant` o `mike wallace`, el primero, jugador de baloncesto, por estar sometido a juicio por una agresión sexual y el segundo, presentador de televisión, por desorden público.

[2] Fue imposible determinar la relevancia de los resultados obtenidos para consultas como `e-find`, `people` o `itt-tech and homework` puesto que, a la vista de la consulta y los resultados, era muy difícil deducir qué buscaba realmente el usuario con esa consulta.

Consulta		Sitio web oficial	Temas distintos	Resultados	Relevancia
olympics	ⓘ	✓	-	8.250.000	100%
hurricane charley			**1**	**598.000**	**100%**
avril lavigne	🗣	✓	-	1.590.000	100%
google	🛠	✓	-	58.400.000	100%
yahoo	🛠	✓	-	123.000.000	100%
ebay	🛠	✓	-	68.400.000	100%
paris hilton	🗣		3	3.420.000	40%
outback jack	❑	✓	-	156.000	100%
mapquest	🛠	✓	-	1.920.000	100%
yahoo.com	🛠	✓	-	123.000.000	100%
james mcgreevey	🗣	✓	-	129.000	100%
kobe bryant	🗣		2	931.000	80%
dan castellaneta	🗣		1	26.500	100%
lindsay lohan	🗣		1	241.000	100%
mike wallace	🗣		4	1.700.000	50%
dawn staley	🗣		1	32.900	100%
britney spears	🗣	✓	-	4.930.000	100%
michael phelps	🗣		3	447.000	80%
nudist+image			**4**	**709.000**	**¿?**
weather	🛠		1	76.800.000	100%
maps	🛠		1	133.000.000	100%
jokes			**1**	**19.800.000**	**100%**
amber frey	🗣		1	77.700	100%
hotmail	🛠	✓	-	21.700.000	100%
thong	**SEX**		**6**	**7.080.000**	**40%**
games			**3**	**274.000.000**	**70%**
jobs	🛠		1	160.000.000	100%
search engines			**1**	**9.380.000**	**100%**
john heffron	🗣	✓	-	36.200	100%
carmen electra	🗣		1	891.000	100%
pamela anderson	🗣	✓	-	2.710.000	100%
camel+toes	**SEX**		**3**	**206.000**	**80%**
hilary duff	🗣	✓	-	811.000	100%
path client			**¿?**	**3.500.000**	**0%**
tattoos			**2**	**6.020.000**	**90%**
nicky hilton	🗣		1	220.000	100%
ashlee simpson	🗣	✓	-	233.000	100%
thongs	**SEX**		**3**	**1.820.000**	**80%**
dictionary			**1**	**38.600.000**	**100%**
home depot	🛠	✓	-	2.630.000	100%
hotmail.com	🛠	✓	-	21.700.000	100%
www.thehun.com	🛠	✓	-	518.000	100%
share jackson			**¿?**	**2.640.000**	**10%**
inuyasha			**1**	**1.270.000**	**100%**
anime			**1**	**42.800.000**	**100%**
ian thorpe	🗣	✓	-	225.000	100%
travel			**1**	**174.000.000**	**100%**
ask jeeves	🛠	✓	-	1.100.000	100%
jessica simpson	🗣	✓	-	1.620.000	100%
ebay.com	🛠	✓	-	68.400.000	100%

Tabla 2. Cincuenta consultas más frecuentes realizadas por internautas de todo el mundo durante el 17 de Agosto de 2004 (Fuente: *Wordtracker*).

Consulta	Sitio web oficial	Temas distintos	Resultados	Relevancia
rapid serial visual presentation		**1**	**34.700**	**100%**
sesshomaru		**1**	**47.900**	**100%**
ffdo		**3**	**2.810**	**60%**
cape plumbago photo		**1**	**450**	**70%**
rosettanet	✓	-	97.900	100%
nike air force ones		2	43.200	¿?
"recetas de ensaladas"		**1**	**6.210**	**70%**
rental car coupons		1	1.560.000	¿?
hemmoroid *(sic)* treatment		**1**	**686**	**100%**
jets pizza		1	42.800	¿?
dee zee running boards	✓	1	8.730	100%
www.adopt a pokemon		**¿?**	**32**	**0%**
how much paper is in one tree?		**1**	**2.200.000**	**70%**
wedding tents		1	181.000	¿?
itt-tech and homework		¿?	1.010	¿?
captree school+west islip	✓	1	14.100	100%
prpcmonitor		**1**	**958**	**100%**
daleville indiana + images		**4**	**3.500**	**20%**
arlington washington public library		**1**	**277.000**	**0%**
fundraising software comparison blackbaud oficial		**1**	**505**	**100%**
handwoven yoga mats		**1**	**899**	**100%**
flood film project		**8**	**283.000**	**30%**
car insurace *(sic)* quotes		**1**	**15.100**	**50%**
"salas surgical group california"		**0**	**0**	**0%**
ravrn simone		**1**	**2**	**0%**
where can i upload roms		**¿?**	**106.000**	**0%**
pcgs + indian cent cameo		**1**	**2.270**	**100%**
what do black baby snakes eat		**1**	**63.300**	**20%**
compile time object		**¿?**	**720.000**	**0%**
ymca swimming lessons- kalamazoo, mi		**1**	**201**	**10%**
combat support flight		2	618.000	¿?
elaine beno		**3**	**836**	**30%**
ancient olympics		**1**	**511.000**	**100%**
methow.com	✓	-	54.100	100%
glenn county, ca		**1**	**891.000**	**100%**
people		¿?	316.000.000	¿?
win ace *(sic)*	✓	2	1.080.000	100%
photgraphed *(sic)* by george smith		**¿?**	**95**	**0%**
"motorcycle tent trailer"		**1**	**2.910**	**0%**
wimbledon	✓	1	2.000.000	100%
required octane		**1**	**90.900**	**20%**
e-find		¿?	53.200	¿?
autos		**1**	**15.500.000**	**100%**
escorps		**2**	**129**	**80%**
nra	✓	7	1.280.000	100%
origin of christian traditions		**1**	**342.000**	**100%**
galaxie 1963 ford power steering valve		**1**	**650**	**20%**
broadband internet tucson		**2**	**36.200**	**80%**
armitage funeral home kearny		**1**	**68**	**100%**
instagate pro vpn		**1**	**1.680**	**100%**

Tabla 3. Cincuenta consultas capturadas en tiempo real el 17 de Agosto de 2004 (Fuente: *Dogpile – Searchspy*).

Puesto que los resultados obtenidos con este pequeño experimento han coincidido con los obtenidos en otros realizados a mayor escala parece razonable no extrapolar sin más los datos obtenidos pero sí argumentar que un porcentaje relativamente elevado de las consultas "típicas" no obtiene resultados relevantes en su primera formulación. Al examinar

una a una las consultas de la muestra parece que algunas podrían reformularse y otras presentan errores ortográficos. No obstante, dada la reticencia de los usuarios a realizar más de una consulta para un mismo problema no parece adecuado exigirles lo primero y, por otro lado, un sistema debería ser robusto frente a errores tipográficos y ortográficos[1] sin recurrir a artefactos complejos.

En resumen, por un motivo u otro, un porcentaje desconocido pero elevado[2] (tal vez entre el 15 y el 20%) de las consultas que se envían a un buscador moderno como *Google* no obtienen ningún resultado positivo entre los diez primeros. Podría aducirse que si los diez documentos más relevantes no satisfacen la consulta entonces ninguno de los restantes lo hará; sin embargo, este argumento es poco sólido.

Por un lado, es necesario recordar que la "relevancia" de los documentos es, en realidad, una medida de su "prestigio" que se obtiene de manera algorítmica basándose, fundamentalmente, en los enlaces que apuntan hacia cada página. Ya se mencionaron algunos problemas de esta técnica señalados por Bharat y Henzinger (1998) –véase página 12. Tales problemas hacen necesario aceptar con reservas el grado de "relevancia" o "irrelevancia" otorgado por un buscador a los distintos documentos de la Web.

Por otro lado, aun en el caso de que todos los enlaces se estableciesen de un modo ideal para los algoritmos de Kleinberg y *PageRank*, habría que seguir desconfiando por una razón muy simple: aunque una página web muy enlazada sea relevante lo contrario, una página poco enlazada es irrelevante, no es necesariamente cierto. Un argumento a favor de esto puede encontrarse en el trabajo de Estelle Broadman (1944) que demostró que el valor de una publicación para un profesional no es directamente proporcional al número de veces en que es citada en otras obras. Recordemos que, de hecho, el índice de impacto de una publicación requiere una normalización del número de citas totales recibidas (Garfield 1972)[3].

Así pues, *Google* y casi todos los buscadores modernos resuelven muy bien aquellas consultas para las que existen una o más páginas "autorizadas" y no obtienen tan buenos resultados cuando no existen tales autoridades. En este último caso, el usuario simplemente recibe una avalancha de información. Para tratar de aliviar esta situación *Google* dispone del

---

[1] Uno de los puntos de interés de la técnica descrita por el autor en esta disertación es, precisamente, su tolerancia al "ruido". No debe confundirse, sin embargo, esta capacidad con la corrección de errores que ofrecen muchos buscadores en la Web. Un sistema "tolerante" aceptaría una consulta como `guttemberg` (*sic*) y ofrecería, de manera transparente, resultados con ese término y otros como `gutemberg`, `guttenberg` e, incluso, `gutenberg`. El hecho de que un documento presente un error tipográfico (o una grafía menos popular) no lo invalida como fuente de información; por ejemplo, alguien que busque información sobre `mao zedong` estará probablemente satisfecho con documentos que mencionen a `mao tse tung`.

[2] Dijimos que no se podían extrapolar los datos puesto que la muestra es muy reducida; no obstante, es posible emitir una suposición razonada. Según Silverstein *et al.* (1998) el 86,4% de las consultas que recibe un buscador (en su caso *Altavista*) se repiten un máximo de tres veces y el 63,7% aparecen una única vez (en ambos casos durante un período de 43 días). Si tomamos estos datos como unas cotas razonables para definir el porcentaje de "consultas típicas" en oposición al de "consultas frecuentes" y damos por bueno el 23% de consultas que no obtienen resultados tendríamos que entre el 15% y el 20% de las consultas que recibe un buscador son resueltas sin dar ningún documento relevante entre los diez primeros.

[3] Podría resultar un problema interesante estudiar posibles modificaciones del algoritmo de Kleinberg o *PageRank* mediante la aplicación de algunas de las técnicas que se emplean para calcular índices de impacto; sin embargo, tampoco es ese el tema de esta disertación.

servicio *Google Answers*[1] ("*Google* responde" o "Respuestas *Google*") que permite a los usuarios hacer preguntas que serán respondidas por otros usuarios expertos tras un pago en metálico. Es decir, una solución "manual" al problema en cuestión.

Sirva esto como un tercer argumento en apoyo de la existencia tangible de una sobrecarga de información en la Web a la espera de una solución automatizada que, como se dijo antes, es el problema que se afronta en este trabajo.

6 Distintas propuestas para luchar contra la sobrecarga de información

La cantidad de información disponible en Internet es enorme (véase página 4) y existe un porcentaje no despreciable (tal vez entre un 15 y un 20%) de consultas que los buscadores no son capaces de resolver satisfactoriamente. Tales consultas obtienen como resultado cientos o miles de documentos sin existir ninguno adecuado entre los primeros aunque es razonable suponer que alguno de los restantes sí es relevante para las necesidades del usuario. El problema radica en encontrar en un conjunto de documentos muy grande unos pocos que sean de interés para el usuario.

A lo largo de los años noventa se realizaron toda una serie de investigaciones sobre este asunto no sólo en la Web sino también en otros servicios como correo electrónico o grupos de *USENET*. Estos trabajos emplearon, de forma independiente o combinada, tres técnicas básicas: agentes *software*, filtrado colaborativo y recomendación por contenidos.

Un **agente** es un elemento *software* capaz de interactuar con su entorno (incluidos otros agentes) para realizar una tarea en representación de un usuario o de otro agente. Los agentes implementan algún tipo de inteligencia artificial que les permite actuar de manera autónoma y determinar las acciones apropiadas para responder a los eventos del entorno.

El **filtrado colaborativo** (Goldberg *et al* 1992) proporciona a un usuario lo que otros individuos similares encontraron de utilidad antes que él. Un ejemplo típico es el servicio de *Amazon*[2] *"Customers who bought this book also bought…"* ("Los clientes que compraron este libro también compraron…")

Por su parte, la **recomendación por contenidos** tiene como objetivo proporcionar documentos similares a un documento de partida y precisa, por tanto, de algún tipo de análisis del texto de los documentos.

A continuación se describirán muy brevemente algunas de las iniciativas más interesantes señalando aspectos innovadores potencialmente aplicables al problema de la sobrecarga de información en la Web así como aspectos débiles de cara a una solución totalmente automática.

Paul E. Baclace (1991 y 1992) utiliza agentes para filtrar la información que recibe un usuario. Dichos agentes evalúan, de manera individual, el interés de los documentos en función del autor y algunas palabras clave. Una vez hecho esto se obtiene una puntuación

---

[1] El servicio *Google* Answers (`http://answers.google.com`) está definido en los términos siguientes: *"El motor de búsqueda de Google es una gran manera de encontrar información en línea. Pero a veces incluso los usuarios experimentados necesitan ayuda para encontrar la respuesta exacta a una pregunta. Google Answers es una forma de conseguir ayuda de expertos en la búsqueda en línea. Al proponer una pregunta usted especifica la cantidad que está dispuesto a pagar por la respuesta y la diligencia con que necesita esa información. Un experto buscará la respuesta y le enviará la información que está buscando, así como enlaces útiles a páginas web sobre el tema. Si usted está satisfecho con la respuesta pagará la cantidad previamente estipulada."*

[2] `http://www.amazon.com`

media para cada documento y aquellos de interés son enviados al usuario que debe evaluarlos. Dicha evaluación permite recompensar a los agentes que evaluaron correctamente el documento y penalizar a los que lo hicieron incorrectamente. Tras una serie de iteraciones el usuario dispone de una población de agentes adaptada a sus intereses. Esta propuesta, aunque interesante, es difícilmente aplicable al problema que nos ocupa por varias razones. En primer lugar, requiere una evaluación explícita por parte del usuario de los documentos calificados como relevantes por los agentes, algo que puede ser inabordable en muchos casos. En segundo lugar, al requerir una serie de iteraciones para obtener un conjunto apto de agentes la técnica es útil para filtrar información acerca de intereses relativamente estables en el tiempo pero no para resolver consultas específicas.

Masahiro Morita y Yoichi Shinoda (1994) describen un experimento que trata el problema de proporcionar artículos interesantes de *USENET* a un grupo de usuarios en función de sus preferencias. El sistema presentado obtiene las valoraciones de manera implícita (a partir de los tiempos de lectura, de las acciones realizadas en el entorno y de las acciones realizadas sobre el texto del artículo) demostrando así que es posible extraer información relevante para el usuario sin necesidad de exigirle ningún esfuerzo consciente. Por otro lado, Morita y Shinoda no utilizan palabras clave (en realidad ideogramas clave) para seleccionar los documentos sino bigramas[1] de palabras.

Pattie Maes (1994) describe una serie de agentes con cometidos similares a los de Baclace (1991 y 1992): filtrar correo y artículos *USENET*, además de recomendar libros o música. Al igual que este último, Maes pretende que el usuario evalúe la calidad de la información que se le ofrece de una manera que se podría calificar de "abiertamente intrusiva". Por ejemplo, el sistema de recomendación musical, *Ringo*, requiere que un usuario evalúe en el momento del registro una lista de 125 artistas para indicar sus preferencias (Shardanand y Maes 1995, p. 211) algo que no parece demasiado razonable.

Menczer, Belew y Willuhn (1995) y Menczer y Belew (1998) describen una técnica similar a la de Baclace (1991 y 1992) aunque con diferencias importantes. En primer lugar, el sistema se emplea para realizar consultas en la Web y no para filtrar información. En segundo, los ecosistemas de agentes se crean para cada consulta individual por lo que no existen ni evolucionan de forma indefinida. Los agentes disponen de una cierta energía que consumen al explorar la Web y pueden recuperar parte de la energía consumida presentando algún documento al usuario que debe valorarlo de manera explícita. Como ya se dijo con anterioridad obligar al usuario a evaluar los resultados no es adecuado, especialmente existiendo formas de obtener una evaluación implícita (Morita y Shinoda 1994). Por otro lado, la evaluación de los prototipos se hace en subgrafos de la Web muy limitados: 116 documentos en el caso de (Menczer, Belew y Willuhn 1995) y 11.000 documentos del sitio web de la *Encyclopaedia Britannica* en (Menczer y Bellew 1998). Además, los resultados de estos experimentos se comparan con las técnicas empleadas por los buscadores de la época que no implementan algoritmos como *PageRank* o similares algo común en los buscadores actuales. Así pues, se trata de una técnica interesante aunque es difícil determinar en qué medida mejoraría los resultados de un buscador moderno.

Henry Lieberman (1995) desarrolló *Letizia*, un agente que asiste al usuario mientras éste navega por la Web. *Letizia* analiza las acciones del usuario sobre los documentos

[1] Un bigrama es una subcadena que contiene dos elementos (palabras o caracteres) y que se obtiene desplazando, elemento a elemento, una "ventana" sobre el texto. La oración anterior, por ejemplo, contendría los siguientes bigramas de palabras: `<Un bigrama>`, `<bigrama es>`, `<es una>`, `<una subcadena>`, etc.

(activar un enlace, grabar o imprimir el documento, etc.) para establecer su interés, determina de forma aproximada el contenido de los documentos extrayendo una serie de palabras clave y, además, explora la Web en segundo plano en búsqueda de documentos similares a los que el usuario considera interesantes. Los documentos valorados como potencialmente interesantes se almacenan en una lista que evoluciona a medida que avanza la exploración del usuario; de tal forma que puede, en cualquier momento, solicitar al agente una recomendación que éste extrae de la lista anterior.

Son varios los aspectos a destacar en esta propuesta: no requiere valoración explícita del usuario, determina un perfil aproximado para el mismo y explora la Web en su representación. Sin embargo, también presenta algunos inconvenientes: el análisis del contenido de los documentos es muy simple y puede conducir a recomendar documentos irrelevantes que coinciden en algunas palabras clave. *Letizia* sólo explora documentos próximos a aquel en que se encuentra el usuario y, además, la experiencia pasada del usuario o de otros usuarios con intereses similares no es tenida en cuenta.

LIRA (Balabanovic, Shoham y Yun 1995) es un agente que permite recomendar diariamente a un usuario un pequeño conjunto de páginas web potencialmente interesantes. Según sus autores, a lo largo del experimento el sistema ofreció en un 50% de los casos mejores resultados que los ofrecidos por un experto humano. Sin embargo, el experimento se hizo con un máximo de 6 usuarios simultáneamente y durante apenas 3 semanas por lo que no pueden considerarse unos datos excesivamente concluyentes. Por otro lado, son varias las críticas que se pueden hacer a *LIRA*: requiere de los usuarios una valoración explícita de los documentos, sólo funciona adecuadamente si el usuario manifiesta un único interés bien definido y, además, emplea extracción de palabras clave como herramienta de análisis de contenidos. Esto último se manifiesta en una limitación reconocida por los propios investigadores: "*Las páginas retornadas por el sistema son a menudo muy similares entre sí, tal y como han señalado muchos de los usuarios.* (Balabanovic, Shoham y Yun 1995, p. 8)"

MUSAG (Goldman, Langer y Rosenschein 1996) es uno de los primeros intentos de abandonar la técnica de coincidencia de palabras clave para la búsqueda de información en la Web. El prototipo utiliza dos agentes, *MUSAG* y *SAg*. El primero tiene como finalidad generar diccionarios "conceptuales" que agrupan las palabras que emplea el usuario en sus consultas con palabras que aparecen en los documentos resultantes. El segundo, *SAg*, emplea estos diccionarios para expandir las consultas. Esta técnica es similar a una de las utilizadas por Salton (1968) en el sistema *SMART* (véase página 3). No obstante, los diccionarios son simples tablas de expresiones asociadas a una palabra y, además, el único criterio de relevancia es la presencia de palabras del diccionario en el documento sin tener en cuenta los posibles intereses o necesidades del usuario.

Fab (Balabanovic y Shoham, 1997) es un sistema de agentes que recomienda páginas web mediante un sistema híbrido que combina colaboración entre usuarios y análisis automático de contenidos. Esta primera versión de *Fab* requería una evaluación explícita de los documentos, posteriormente sería modificado en la línea de (Morita y Shinoda 1994) para obtener valoraciones implícitas:

> *En escenarios típicos, los usuarios proporcionan feedback explícito sólo a regañadientes [...] por tanto, no es razonable imponer una carga extra a usuarios que ya intentan reducir su sobrecarga de información. Por tanto, el primer objetivo es aprender a recomendar documentos apropiados utilizando solamente feedback implícito.* (Balabanovic 1998, p. 6)

Es necesario indicar que en las pruebas que Balabanovic realizó de su sistema se emplearon 1.600 artículos de prensa cubriendo un período de dos semanas, documentos, quizás, demasiado homogéneos en cuanto a su estructura y muy diferentes de la mayor parte de páginas web existentes.

GroupLens (Konstan *et al.* 1997) describe un sistema que demuestra que la utilización del tiempo de lectura de un documento como sistema de evaluación implícita permite obtener recomendaciones similares a las producidas empleando valoración explícita.

Siteseer (Rucker y Marcos 1997) es un proyecto sencillo pero que señala un par de puntos interesantes. El sistema tomaba los *bookmarks* (páginas favoritas) de un usuario y su estructuración como un indicativo de sus intereses y las relaciones semánticas que establecía entre los mismos. Para realizar recomendaciones, se comparaban los intereses de cada usuario con los del resto y se le aconsejaba visitar documentos "favoritos" de otros usuarios que no estuviesen en su lista[1].

AntWorld (Kantor *et al.* 2000) es un proyecto que trata de ayudar a los usuarios a encontrar la información que buscan aprovechando la experiencia y valoraciones de anteriores usuarios del sistema. Como la mayoría de las propuestas revisadas, los desarrolladores de *AntWorld* piensan que la valoración de los documentos debe ser activa por parte del usuario y se muestran escépticos sobre la obtención pasiva de dicha valoración.

En resumen, para poder ofrecer a un usuario unos pocos documentos seleccionados de un conjunto muy grande es inevitable que el propio usuario u otros usuarios con intereses similares los hayan "evaluado". No obstante, no es necesario que la evaluación de los documentos sea explícita (por ejemplo, otorgando una calificación) puesto que el comportamiento del usuario al actuar sobre el documento proporciona indicios sobre el grado de interés del mismo (Morita y Shinoda 1994), (Balabanovic 1998) o (Jansen y Spink 2003).

Por otro lado, también es imprescindible realizar un análisis de los contenidos a fin de determinar qué documentos y/o perfiles de usuarios (páginas favoritas, historial de visitas, consultas realizadas, combinaciones de lo anterior, etc.) son similares y en qué grado. Morita y Shinoda (1994) señalaron que es posible emplear técnicas sencillas (en su caso bigramas de ideogramas) para comparar documentos con mejores resultados que empleando únicamente palabras clave.

7 La Web Semántica

Paralelamente al desarrollo de técnicas como las de Kleinberg o *Google* para localizar documentos en la Web y al mismo tiempo en que se buscaban soluciones al problema de la sobrecarga de información en Internet, Tim Berners-Lee (1998) esbozaba el concepto de **Web Semántica** que, junto con James Hendler y Ora Lassila, refinaría posteriormente (Berners-Lee, Hendler y Lassila 2001).

Simplificando enormemente puede decirse que el objetivo básico de la Web Semántica es permitir que agentes *software* sean capaces de "consumir" documentos disponibles en la Web para inferir nuevo conocimiento. Para ello los documentos deberían construirse empleando lenguajes "semánticos" que permitirían no sólo anotar

[1] Una iniciativa similar a la que ahora desarrolla *del.icio.us* que se define como *"un gestor social de enlaces favoritos"* (`http://del.icio.us`).

metainformación sino también especificar las relaciones existentes entre los metadatos. El *quid* de la cuestión radica en la forma de construir las etiquetas semánticas de los nuevos lenguajes e indicar las relaciones entre las mismas (véase Fig. 1 y Fig. 2). Para realizar esta labor se ha optado por la utilización de **ontologías**.

```
<INSTANCE KEY="http://www.cs.umd.edu/users/hendler/">
  <USE-ONTOLOGY ID="cs-dept-ontology" VERSION="1.0" PREFIX="cs"
  URL="http://www.cs.umd.edu/projects/plus/SHOE/cs.html" />

  <CATEGORY NAME="cs.Professor" FOR="http://www.cs.umd.edu/users/hendler/"/>

  <RELATION NAME="cs.member">
    <ARG POS=1 VALUE="http://www.cs.umd.edu/projects/plus/">
    <ARG POS=2 VALUE="http://www.cs.umd.edu/users/hendler/">
  </RELATION>

  <RELATION NAME="cs.name">
    <ARG POS=2 VALUE="Dr. James Hendler">
  </RELATION>

  <RELATION NAME="cs.doctoralDegreeFrom">
    <ARG POS=1 VALUE="http://www.cs.umd.edu/users/hendler/">
    <ARG POS=2 VALUE="http://www.brown.edu">
  </RELATION>

  <RELATION NAME="cs.emailAddress">
    <ARG POS=2 VALUE="hendler@cs.umd.edu">
  </RELATION>

  <RELATION NAME="cs.head">
   <ARG POS=1 VALUE="http://www.cs.umd.edu/projects/plus/">
   <ARG POS=2 VALUE="http://www.cs.umd.edu/users/hendler/">
  </RELATION>
</INSTANCE>
```

Fig. 1 Código *SHOE* (véase página 25) **utilizado para etiquetar una página *HTML*.**

Este código parte de una ontología (véase Fig. 2) que describe departamentos universitarios de informática. Indica que el documento (la página *HTML*) hace referencia al profesor James Hendler que es doctor por la Universidad de Brown y director de una organización cuya información está disponible en `http://www.cs.umd.edu/projects/plus/`.

El uso del término "ontología" no está exento de polémica (Soergel 1999) o (Bates 2002) debida, en parte, al origen filosófico[1] del mismo. Sin embargo, la definición de ontología aplicable al campo de la Web Semántica tiene poco que ver con la filosofía:

> *Una ontología es la especificación de una conceptualización. Esto es, una descripción de los conceptos y relaciones que pueden existir para un agente o una comunidad de agentes* (Gruber 1993).

Según Berners-Lee, Hendler y Lassila (2001, p. 4) una ontología es:

> *Un documento o fichero que define formalmente las relaciones entre términos. Una ontología típica para la Web consta de una taxonomía y de un conjunto de reglas de inferencia.*

Con anterioridad o simultáneamente a la propuesta de Berners-Lee (1998) para la Web Semántica se estaban realizando una serie de trabajos que tenían como objetivo desarrollar lenguajes que permitiesen definir tales ontologías y utilizarlas para etiquetar documentos en la Web, lo que podríamos denominar "pre-Web-Semántica". Cabe destacar los proyectos *SHOE* (Luke, Spector y Rager 1996), *WebKB* (Craven *et al.* 1998) y *Ontobroker/On2broker* (Fensel *et al.* 1998) y (Fensel *et al.* 1999), respectivamente.

[1] Según el Diccionario de la Lengua Española (*RAE* 2001) la ontología es la "parte de la metafísica que trata del ser en general y de sus propiedades trascendentales."

```
<!-- The ontology declarations start here. We begin by creating the ontology -->

<ONTOLOGY ID="cs-dept-ontology" VERSION="1.0" DESCRIPTION="An example ontology for
computer science academic department">

<!-- Now we declare that the ontology will be borrowing elements from
     the base-ontology.  -->

<USE-ONTOLOGY ID="base-ontology" VERSION="1.0" PREFIX="base"
URL="http://www.cs.umd.edu/projects/plus/SHOE/onts/base1.0.html">

<!-- Here we declare the categories in this ontology -->

<DEF-CATEGORY NAME="Person" ISA="base.SHOEEntity" SHORT="person">
<DEF-CATEGORY NAME="Worker" ISA="Person" SHORT="worker">
<DEF-CATEGORY NAME="Faculty" ISA="Worker" SHORT="faculty member">
<DEF-CATEGORY NAME="Professor" ISA="Faculty" SHORT="professor">
…
<DEF-CATEGORY NAME="Organization" ISA="base.SHOEEntity" SHORT="organization">
…
<DEF-CATEGORY NAME="ResearchGroup" ISA="Organization" SHORT="research group">
…

<!-- Here we declare the relations in the ontology -->

<DEF-RELATION NAME="emailAddress" SHORT="can be reached at">
  <DEF-ARG POS=1 TYPE="Person">
  <DEF-ARG POS=2 TYPE=".STRING" SHORT="email address">
</DEF-RELATION>

<DEF-RELATION NAME="head" SHORT="is headed by">
  <DEF-ARG POS=1 TYPE="Organization">
  <DEF-ARG POS=2 TYPE="Person">
</DEF-RELATION>
…
<DEF-RELATION NAME="doctoralDegreeFrom" SHORT="has a doctoral degree from">
  <DEF-ARG POS=1 TYPE="Person">
  <DEF-ARG POS=2 TYPE="University">
</DEF-RELATION>
…
<DEF-RELATION NAME="member" SHORT="has as a member">
  <DEF-ARG POS=1 TYPE="Organization">
  <DEF-ARG POS=2 TYPE="Person" SHORT="member">
</DEF-RELATION>

<!-- Here we declare some example inferences which might be useful to agents.  -->

<DEF-INFERENCE DESCRIPTION="Transitivity of Suborganizations.  If subOrganization(x,y)
and subOrganization(y,z) then subOrganization(x,z)">
  <INF-IF>
    <RELATION NAME="subOrganization">
      <ARG POS=FROM VALUE="x" VAR>
      <ARG POS=TO VALUE="y" VAR>
    </RELATION>
    <RELATION NAME="subOrganization">
      <ARG POS=FROM VALUE="y" VAR>
      <ARG POS=TO VALUE="z" VAR>
    </RELATION>
  </INF-IF>
  <INF-THEN>
    <RELATION NAME="subOrganization">
      <ARG POS=FROM VALUE="x" VAR>
      <ARG POS=TO VALUE="z" VAR>
    </RELATION>
  </INF-THEN>
</DEF-INFERENCE>
…

<!-- The end of the ontology marked here -->

</ONTOLOGY>
```

Fig. 2 Ontología *SHOE* (véase página 25) **que describe un departamento de informática.**

Obsérvese cómo la ontología describe categorías (p.ej. persona, trabajador, profesor, organización, grupo de investigación, etc.), relaciones (p.ej. dirección de correo, doctorado por, miembro de, etc.) y ofrece ejemplos de inferencia de conocimiento.

SHOE (Luke, Spector y Rager 1996) es una de las primeras iniciativas destinadas a proporcionar un lenguaje de marcado semántico. Se trata de una extensión del lenguaje *HTML* que permite desarrollar ontologías (véase Fig. 2) y utilizar las clases y relaciones definidas en una o más de esas ontologías para marcar zonas específicas de un documento *HTML* (véase Fig. 1). Luke *et al.* describen asimismo una herramienta, *Exposé*, que explora la Web en busca de páginas anotadas con *SHOE* y almacena los asertos que encuentra en una base de conocimiento que puede utilizarse posteriormente para realizar consultas.

WebKB (Craven *et al.* 1998) tenía como objetivo construir, de forma automática, una base de conocimiento que reflejase el contenido de la Web de una forma inteligible para una máquina. Para lograr esto el sistema debía recibir una ontología que describiese las clases y relaciones, así como un conjunto de documentos, etiquetados sobre la base de dicha ontología, que servirían como conjunto de entrenamiento. Así, tras un período de entrenamiento adecuado, el sistema sería capaz de procesar documentos *HTML* y producir documentos marcados semánticamente de acuerdo a la ontología de partida.

Ontobroker (Fensel *et al.* 1998) fue una iniciativa muy similar a *SHOE* puesto que proponía una serie de herramientas para definir ontologías, etiquetar documentos basándose en dichas ontologías y realizar consultas e inferencia sobre una base de conocimiento. Posteriormente evolucionaría hacia *On2broker* (Fensel *et al.* 1999) cuya principal novedad fue la utilización de tecnologías como *XML*[1] o *RDF*[2].

XML y *RDF* constituyen las bases sobre las que comenzar a construir la Web Semántica (Berners-Lee, Hendler y Lassila 2001, p. 3) puesto que el primero posibilita la construcción de nuevos lenguajes de etiquetas, por ejemplo *RDF* que, a su vez, permite expresar asertos. Sin embargo, son necesarias toda una serie de capas encima de *RDF* para desarrollar finalmente la Web Semántica.

Por ejemplo, aunque *RDF* permite dar valores a las distintas propiedades de diferentes recursos no dispone de mecanismos para describir esas propiedades ni para describir las relaciones entre las propiedades y otros recursos. Para ello es necesario un lenguaje que permita definir vocabularios *RDF*. Dicho lenguaje, construido mediante *RDF*, es *RDF Schema* o *RDF(S)* (Brickley y Guha 2004). Este lenguaje define clases y propiedades que permiten, a su vez, describir nuevas clases, propiedades y recursos.

Sin embargo, tampoco *RDF* ni *RDF Schema* son capaces por sí solos de modelar ontologías, razón por lo que comienzan a desarrollarse lenguajes para este fin análogos a los definidos durante la fase pre-Web-Semántica con la diferencia de que los nuevos lenguajes se construyen sobre el estándar *RDF(S)*. Ejemplos de estas extensiones ontológicas para

---

[1] *XML – eXtensible Markup Language* (Lenguaje de Etiquetado Extensible) es una recomendación del Consorcio W3 que permite crear lenguajes de etiquetado de propósito específico (vocabularios *XML*) `<http://www.w3.org/XML/>`.

[2] *RDF – Resource Description Framework* (Marco para la Descripción de Recursos) es la especificación de un modelo de metadatos que realiza descripciones de recursos mediante sentencias que combinan un objeto, una propiedad y un valor para dicha propiedad, todo ello serializado mediante *XML* `<http://www.w3.org/RDF/>`. A continuación se muestra la sentencia "*Daniel Gayo Avello es el autor de* `http://www.di.uniovi.es/~dani/`":

```
<rdf:Description about='http://www.di.uniovi.es/~dani/'>
<Author>Daniel Gayo Avello</Author>
</rdf:Description>
```

RDF Schema son las desarrolladas por Staab *et al.* (2000) y Horrocks *et al.* (2000) que definen *OIL*[1] o por McGuinness *et al* (2000) con *DAML-ONT*[2].

Posteriormente *DAML-ONT* y *OIL* convergieron en el lenguaje *DAML+OIL* (van Harmelen, Patel-Schneider y Horrocks 2001) que terminaría evolucionando hacia *OWL*[3] (Bechhofer *et al.* 2004) una recomendación del Consorcio W3 y, por tanto, el estándar para la construcción de ontologías.

8 Consultas en la Web Semántica

En estos momentos existen toda una serie de tecnologías estandarizadas por el Consorcio W3 que permiten construir parte de la "pila" de la Web Semántica (véase Fig. 3). Así, se utiliza *XML* para desarrollar vocabularios como *RDF*, el cual permite expresar asertos acerca de recursos disponibles en la Web. Éste, a su vez, es la base para construir *RDF(S)* que posibilita la creación de nuevos vocabularios *RDF* y, por tanto, la creación de un lenguaje como *OWL* para definir ontologías. Ontologías que, a su vez, permiten etiquetar los documentos de la Web Semántica.

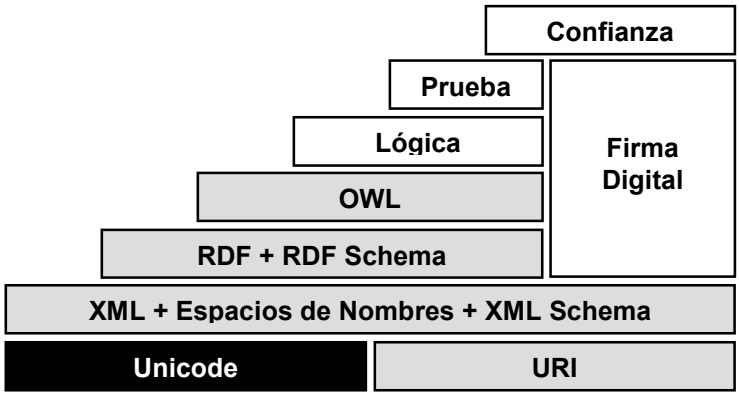


Fig. 3 Pila de la Web Semántica.

Se muestran sombreados aquellas capas de la Web Semántica para las que ya existe un estándar. En gris claro se representan los estándares propuestos por el Consorcio W3.

Sin embargo, aún quedan por desarrollar una serie de capas de esa "pila" y, quizás, una de las más urgentes sea la capa de consulta[4]. No obstante, ya se han investigado toda una serie de lenguajes entre los que cabe destacar *Metalog*, *SquishQL*/*RDQL* o *RQL*/*SeRQL*.

Metalog (Marchiori y Saarela 1998, 1999a y 1999b) fue el primer sistema en añadir una capa de lógica y consulta sobre *RDF* permitiendo inferir conocimiento nuevo a partir de los metadatos disponibles. Para facilitar su uso emplea un lenguaje "pseudo-natural" (*LPN*), es decir, una versión simplificada y controlada del inglés que permite expresar hechos, reglas y consultas que se aplicarán sobre metainformación *RDF*. Estas acciones pueden expresarse no sólo en *LPN* sino también en *RDF* o mediante algún otro lenguaje de programación lógica.

---

[1] *Ontology Inference Layer*, Capa de Inferencia Ontológica.

[2] Lenguaje ontológico del programa *DARPA Agent Markup Language* (Lenguaje de Etiquetado para Agentes DARPA).

[3] *Web Ontology Language* (Lenguaje Ontológico para la Web).

[4] Más bien inferencia (Guha *et al.* 1998).

SquishQL (Brickley y Miller 2000) (Miller, Seaborne y Reggiori 2002) es un lenguaje similar a *SQL* para realizar consultas sobre *RDF*. *RDQL* (Seaborne 2004) puede considerarse una evolución de *SquishQL* y ha sido propuesto por *Hewlett-Packard* al Consorcio W3 como un posible lenguaje de consulta para *RDF*.

RQL (Karvounarakis et al. 2001) es un lenguaje funcional, integrado dentro de *Sesame*[1], que permite consultar datos *RDF* y *RDF(S)*. *Sesame* tiene como objetivo ofrecer una plataforma estable para almacenar, consultar, manipular y administrar ontologías y metadatos expresados no sólo en *RDF* o *RDF(S)* sino también en *OWL*. En la actualidad, el lenguaje *SeRQL* (*Aduna B.V.* y *Sirma AI Ltd.* 2002-2004) ha sustituido a *RQL*.

Estos lenguajes probablemente tendrán una gran influencia en un futuro estándar W3C para un lenguaje de consulta sobre *RDF*, actualmente[2] en estudio por parte del *RDF Data Access Working Group* (Grupo de Trabajo para Acceso a Datos RDF). Después de todo Janne Saarela (*Metalog*), Alberto Reggiori (*SquishQL*), Andy Seaborne (*SquishQL* y *RDQL*) o James Hendler (*SHOE* y *DAML-ONT*) son algunos de los miembros de este grupo. Hasta el momento se han especificado casos de uso, requisitos y objetivos para el lenguaje de consulta (Clark 2004) y se ha empezado a esbozar el lenguaje *BRQL*[3] (Prud'hommeaux y Seaborne 2004) que permitirá seleccionar información, extraer subgrafos *RDF* y construir nuevos grafos *RDF* a partir del resultado de una consulta. Por el momento no se contempla utilizar el lenguaje de consulta sobre *RDF(S)* ni *OWL*.

```
SELECT ?title
PREFIX dc: <http://purl.org/dc/elements/1.1/>
WHERE { <http://example.org/book/book1> dc:title ?title . }
```

Fig. 4 Consulta BRQL para determinar el título de un libro.

Si *BRQL*, o un lenguaje similar, se convierte finalmente en un primer estándar para consultas en la Web Semántica (algo muy probable) se abrirán toda una serie de posibilidades en campos como agregación de contenidos, transporte, sistemas de producción, turismo, gestión de información personal, pruebas de *software*, comercio electrónico, etc. Los casos de uso planteados (Clark 2004) auguran un lenguaje expresivo y potente, aunque para resolver consultas fundamentalmente "metasemánticas", por ejemplo[4]:

- Encontrar la dirección de correo de Jonhny Lee Outlaw.
- Encontrar en la web de un proveedor información sobre el repuesto de una pieza así como la lista de piezas que deben ser sustituidas junto con la defectuosa.
- Recibir puntualmente información sobre libros, películas y música que cumplan unos criterios de título, precio y autor.
- Grabar todos los programas de televisión sobre el jugador de béisbol Ichiro.

La utilidad de una tecnología que permita resolver necesidades como las anteriores está fuera de toda duda; sin embargo, el problema que se afronta en este trabajo no es ese sino la forma de resolver en la Web de manera adecuada y automática consultas

[1] `http://www.openrdf.org`

[2] Agosto de 2004.

[3] *Bristol RDF Query Language.*

[4] Se han indicado las "necesidades de información" que se podrían resolver con un lenguaje como *BRQL* no las consultas expresadas en dicho lenguaje.

informativas[1] mucho más abiertas y ambiguas, formuladas en cualquier lenguaje natural[2] (tal vez con errores tipográficos, ortográficos o gramaticales) y susceptibles de sobrecargar de información al usuario. En definitiva, consultas como las siguientes[3]:

- *history and cultural Bengal* (historia y Bengal cultural).
- *acute predictors of aspiration pneumonia: how important is dysphagia?* (predictores adecuados para la neumonía por aspiración: ¿cuán importante es la disfagia?)
- *degenerative disk disease* (enfermedad degenerativa de disco). En el contexto médico es más común la forma *"disc"* que *"disk"*.
- *muscel* (sic) *aches during pregnancy* (dolores *muscualres* durante el embarazo). Consulta con error tipográfico.

9 La Web Cooperativa

Más de cuatro décadas transcurrieron entre la descripción de Vannevar Bush del dispositivo "memex" (Bush 1945) y el desarrollo de la Web, el sistema que tal vez se haya aproximado más a sus ideas. Durante ese tiempo se fueron solventando las "dificultades técnicas de todo tipo" que Bush auguraba y se ha llegado a la situación actual en la que centenares de miles de millones de documentos[4] están, en principio, a un *clic* de distancia de millones de usuarios.

No obstante, la realidad es muy distinta. A no ser que los usuarios conozcan la dirección de los documentos estos son inalcanzables puesto que la Web no dispone, por sí misma, de ningún mecanismo de recuperación de documentos. Por esa razón se han desarrollado sistemas de búsqueda capaces de proporcionar a los usuarios direcciones de páginas web en respuesta a sus consultas. Sin embargo, debido al tamaño de la Web las respuestas son, en general, demasiado numerosas y surge un problema de "sobrecarga de información".

A lo largo de los apartados anteriores se ha expuesto la existencia de dicho problema no sólo en la Web sino también en otros servicios de Internet. Se han estudiado con cierto detalle técnicas interesantes para solucionarlo como la recuperación y filtrado de información, la evaluación explícita o implícita de la relevancia de un documento por parte de los usuarios, las técnicas empleadas para explorar la Web a fin de localizar documentos desconocidos, los métodos para determinar el "prestigio" de un sitio web de modo análogo a como se calcula el "impacto" de una publicación científica así como la futura evolución de la Web hacia la Web Semántica.

---

[1] *"El propósito de las consultas informativas es encontrar información que se supone está disponible en la Web de forma estática. No se prevé más interacción que la lectura. Por forma estática se entiende que el documento no se crea en respuesta a la consulta."* (Broder 2002, p. 5)

[2] Aunque las técnicas que se describirán más adelante podrían ser aplicadas, en teoría, a idiomas ideográficos con resultados análogos a los obtenidos con idiomas alfabéticos, el autor se ha centrado en los segundos.

[3] Fuente: *Dogpile – Searchspy* <http://www.dogpile.com/info.dogpl/searchspy>

[4] Si se acepta que la Web Oculta tiene un tamaño 50 veces superior al de la Web superficial (Aguillo 2002), (Bergman 2001) y se toma el número de páginas indexadas por *Google* como cota inferior del tamaño de la segunda.

De este modo se ha podido delimitar mejor el problema que se afronta en este trabajo:

> *"La sobrecarga de información que experimentan los usuarios al tratar de resolver en la Web consultas informativas formuladas en lenguaje natural de manera tal vez ambigua y, en ocasiones, con errores tipográficos, ortográficos o gramaticales."*

La **Web Cooperativa**[1] (Gayo Avello y Álvarez Gutiérrez 2002) es una propuesta del autor para solucionar ese problema que se sustenta en los siguientes puntos:

- La utilización de conceptos, generados automáticamente, como una alternativa intermedia entre las ontologías y las palabras clave.
- La clasificación de documentos en una taxonomía a partir de tales conceptos.
- La cooperación entre usuarios, en realidad, entre agentes que actúan en representación de los usuarios y que no requieren su participación explícita.

9.1 Conceptos frente a palabras clave

La recuperación de información mediante palabras clave utilizada por los actuales motores de búsqueda plantea dos graves problemas: una tasa de recuperación excesiva y una precisión relativamente baja. La utilización de ontologías puede mejorar la precisión en algunos casos. Sin embargo, desarrollar ontologías que den soporte a cualquier consulta concebible en la Web supondría un esfuerzo inabordable[2].

Existe, sin embargo, una posibilidad intermedia: la utilización de conceptos. Un concepto sería una entidad más abstracta y, por tanto, con mayor carga semántica que una palabra clave. No obstante, no requeriría "artefactos" complejos como lenguajes ontológicos o sistemas de inferencia. Un concepto podría ser considerado como un grupo de palabras con un significado similar, o relacionado, dentro de un ámbito determinado ignorando tiempo, género y número[3]. Por ejemplo, en un área del conocimiento podría existir el concepto (`ordenador`, `máquina`, `servidor`) mientras que en otro existiría (`actor`, `actriz`, `artista`, `celebridad`, `estrella`).

Los conceptos, así entendidos, serán útiles si permiten proporcionar semántica de forma análoga a las ontologías y, simultáneamente, son generados y procesados automáticamente como las palabras clave. El autor tiene puestas grandes esperanzas en las

---

[1] El resto de este apartado y el siguiente se han elaborado a partir de los capítulos tercero y cuarto del trabajo de investigación defendido por el autor en 2002. El primero de ellos se corresponde con un artículo presentado en *COMPSAC* (Gayo Avello y Álvarez Gutiérrez 2002) y en el segundo se aclaran algunos puntos en respuesta a comentarios de los revisores.

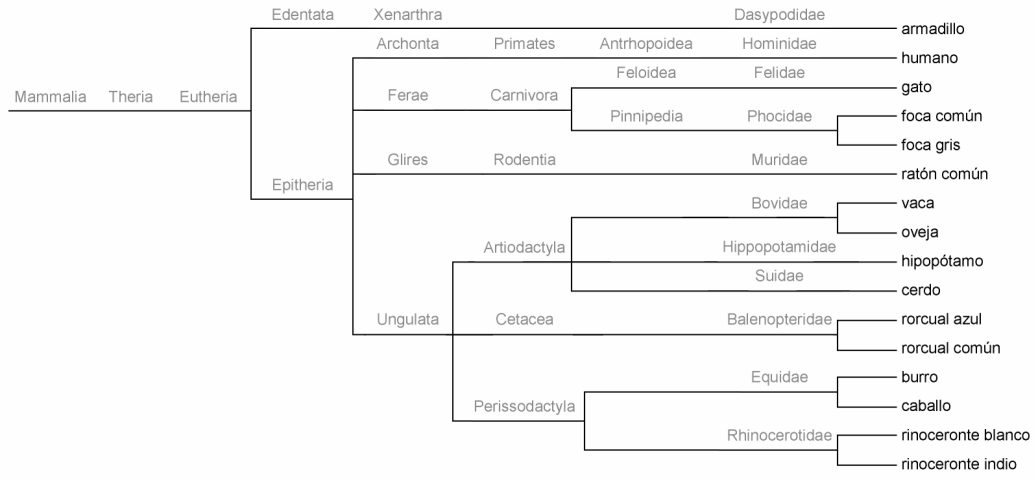
[2] El autor no es el único en sugerir la necesidad de un complemento para la Web Semántica que opere sobre la Web actual. Flake, Pennock y Fain (2003) afirman: *"Muchos han defendido la Web Semántica como un medio para mejorar la recuperación de información en la Web [argumentando que], en su forma actual, no resulta adecuada para el procesamiento automático puesto que la información no está estructurada. […] En la Web Semántica los autores utilizarán un lenguaje para anotar con etiquetas semánticas los datos. […] Resulta sencillo prever el etiquetado [semántico] implícito de catálogos de productos pero podría ser desalentador anotar semánticamente largos pasajes de texto –por ejemplo, artículos de revista. […] Un escenario complementario prevé algoritmos suficientemente inteligentes como para inferir semántica de la Web actual, no estructurada pero auto-organizada, sin ayuda de etiquetas semánticas. […] Los usuarios se beneficiarán más si el trabajo para la creación de la Web Semántica se realiza en paralelo al desarrollo de herramientas para el análisis de datos en la Web auto-organizada."*

[3] Estos conceptos serían similares a los *synsets* (conjuntos de sinónimos) empleados por *WordNet* <`http://wordnet.princeton.edu`>. Los *synsets* se definen como conjuntos de palabras intercambiables en algún contexto.

técnicas de Semántica Latente[1] (Foltz 1990) o de indexación de conceptos (Karypis y Han 2000). En la siguiente sección se examinará la forma en que es posible obtener semántica a partir de conceptos sin emplear ningún soporte ontológico.

9.2 Taxonomías de documentos

Para dotar de significado a un documento, la Web Semántica precisa una ontología que defina una serie de términos y relaciones entre los mismos. Dichos términos son utilizados para etiquetar diferentes partes del documento proporcionando así un "marcado semántico". La Web Cooperativa, por su parte, pretende utilizar el texto completo del documento, sin ningún tipo de etiquetado, como fuente de semántica. ¿Es esto posible sin "comprender" el significado del texto? A lo largo de esta sección se presentará una forma de procesar lenguaje natural para obtener, de manera totalmente automática, una clasificación conceptual de documentos.



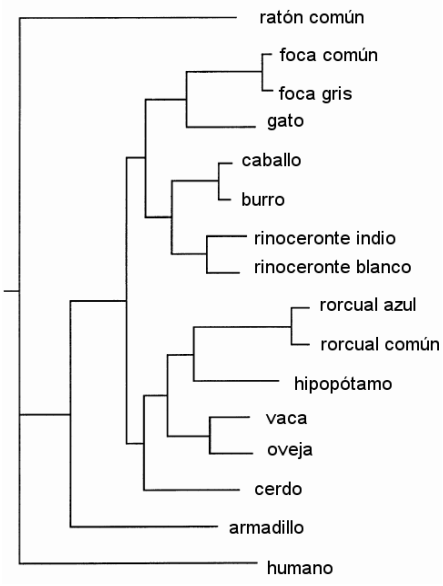


Fig. 5 Clasificación tradicional y clasificación automática basándose en el ADN mitocondrial (Ursing y Arnason 1998).

El método automático ha agrupado correctamente especies muy similares (foca común y gris, rorcual común y azul, vaca y oveja) y ha establecido categorías de un nivel superior (las dos especies de foca junto con el gato, o el caballo y el burro junto a las dos especies de rinoceronte). El método automático ha situado una especie, hipopótamo, en una categoría más próxima a las dos especies de rorcual de lo que aparece en la clasificación tradicional (fenotípica). Por otro lado, especies muy distintas como el ratón, el armadillo o los humanos no han sido agrupados en ninguna categoría próxima al resto de especies, lo cual es correcto. El método genético, sin embargo, no ha podido determinar que el armadillo pertenece a una categoría propia. Este hecho no es importante de cara a la utilización de estos algoritmos como sistema de clasificación semántica de documentos puesto que lo fundamental es establecer relaciones entre documentos similares y separar los documentos diferentes más que determinar el grado exacto de estas diferencias.

[1] Foltz y Dumais (1992) describen una experiencia en la que se combinan dos técnicas diferentes para describir los intereses de un grupo de usuarios (palabras clave y valoración de documentos) y dos técnicas de recuperación de información (búsqueda por palabras clave y semántica latente); la combinación que mejores predicciones produjo fue la semántica latente combinada con valoración de documentos.

Un documento puede considerarse como un individuo de una población. Entre los seres vivos un individuo está definido por su genoma, el cual se compone de cromosomas que se dividen en genes construidos a partir de bases genéticas. De forma similar, los documentos están compuestos por pasajes que se dividen en sentencias construidas mediante conceptos. Siguiendo esta analogía se puede conjeturar que dos documentos estarán semánticamente relacionados si sus respectivos "genomas" son similares y que grandes diferencias entre dichos "genomas" implicarán una relación semántica baja.

El autor considera que esta analogía puede ser puesta en práctica y que es posible adaptar algoritmos empleados en biología computacional al campo de la clasificación de documentos. Simplificando mucho, estos algoritmos se limitan a trabajar con largas cadenas de caracteres que representan fragmentos del genoma de individuos de la misma o de distintas especies. Individuos o especies similares muestran similitudes en sus códigos genéticos de tal forma que es posible mostrar la relación existente entre individuos y especies en taxonomías o dendrogramas[1] sin la necesidad de conocer, o lo que es lo mismo, comprender, la función de cada gen.

Estos dendrogramas permiten, en cierto modo, agrupar a las distintas especies en "categorías"; dichas "categorías" aportan información muy útil para comprender la evolución de las especies y, en muchas ocasiones, confirman (y en otras refutan) el sistema de clasificación de las especies clásico, basado en el sistema linneano.

Este sistema establece los grupos taxonómicos sobre la base de características observables en los seres vivos, es decir, su fenotipo. Los dendrogramas, sin embargo, establecen los distintos grupos basándose en el genotipo de las especies. El fenotipo depende del genotipo pero también está influenciado por el ambiente y por la interacción entre éste y el genotipo. Por esta razón, categorías obtenidas de forma automática a partir de la bioquímica de las especies pueden parecerse extraordinariamente a aquellas otras establecidas mediante un criterio de clasificación consciente e inteligente[2].

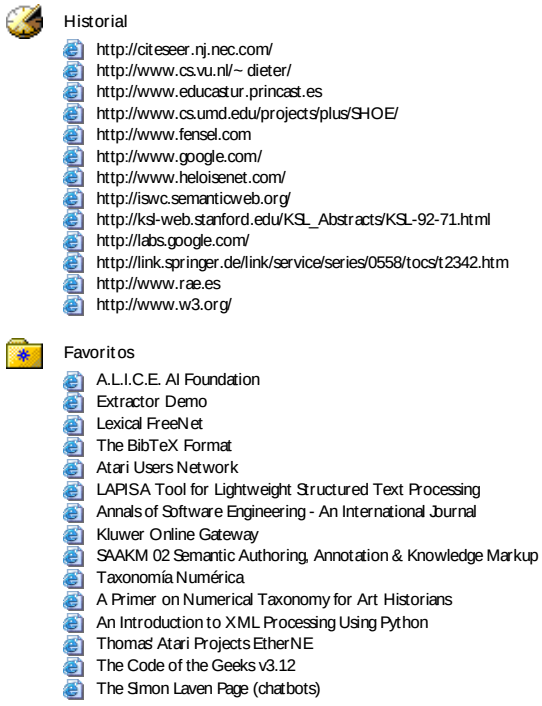


Fig. 6 Historial y lista de favoritos de un usuario.

[1] Un dendrograma es una representación gráfica de un proceso de agrupamiento que muestra las relaciones entre una serie de grupos. Puede verse un dendrograma como un árbol jerárquico, donde los grupos de la misma rama están más relacionados entre sí que con grupos de otras ramas (véase Fig. 5).

[2] Chakrabarti *et al.* (1998b) también plantearon la posibilidad de construir taxonomías de páginas relevantes de forma automática, los resultados que obtuvieron con su sistema *CLEVER* mostraban que las técnicas que empleaban proporcionaban mejores resultados que un directorio generado de forma semi-automática como *Yahoo!*. Sin embargo, el autor cree que es posible generar taxonomías para cualquier documento (no sólo los más relevantes) además de poder emplearse mejores indicadores de la relevancia que los empleados en *CLEVER*.

De forma análoga, los documentos podrían ser clasificados automáticamente en dendrogramas en función de las similitudes encontradas en sus respectivos "genomas conceptuales". La importancia de semejante sistema de clasificación radica en el hecho de que proporcionaría información semántica (similitudes a un nivel conceptual entre distintos documentos o entre documentos y consultas de usuario) sin utilizar ningún tipo de información semántica durante el proceso de clasificación. De hecho, debería ser capaz de agrupar documentos en categorías análogas a las que establecería un ser humano independientemente de la naturaleza del documento y del idioma en que el documento estuviera escrito.

9.3 Colaboración entre usuarios

Ya se ha dicho con anterioridad que la Web actual no permite sacar el máximo provecho al conocimiento experimental que obtienen los usuarios al explorarla. También se han estudiado algunas iniciativas de filtrado y recomendación de información que permitían la participación de los usuarios pero obligaban a estos a valorar documentos o proporcionar información de forma explícita. La Web Cooperativa pretende utilizar estas experiencias para extraer semántica de las mismas de forma no intrusiva y transparente para el usuario. Para ello cada usuario de la Web Cooperativa dispondría de un agente con dos objetivos: aprender de su "maestro" y recuperar información para él.

Fig. 7 Perfil de usuario extraído de los documentos anteriores (véase Fig. 6).

A partir de los documentos presentes en el historial de navegación y la lista de favoritos de un usuario será posible determinar sus principales temas de interés. Estos temas de interés configurarán un perfil que, con fines únicamente ilustrativos, se representa aquí como una "bolsa" de conceptos asociada a documentos representativos. Los distintos temas supondrán un porcentaje determinado del perfil del usuario y cada tema, a su vez, podrá matizar los conceptos que lo constituyen (representado aquí mediante una escala de gris donde los tonos oscuros indican conceptos importantes para el usuario y los claros señalan los menos relevantes).

9.3.1 Aprendizaje de los intereses del maestro

Para alcanzar este objetivo el agente debe desarrollar un perfil que describa de forma precisa los intereses del usuario. Esta descripción se haría mediante los conceptos anteriormente descritos y se construiría a partir de los documentos que el usuario almacena en su equipo, visita con frecuencia, añade a su lista de favoritos, etc. Todo ello sin intervención explícita del usuario.

Una vez un usuario es vinculado a un perfil es posible utilizar esta información para dar una semántica a los documentos de la Web que no es implícita a los mismos sino que depende de los usuarios. Ni la Web actual ni la Web Semántica tienen en cuenta la "utilidad" de los documentos. Los documentos son buscados y procesados por la utilidad que los usuarios esperan obtener de ellos. La utilidad de un documento no reside en sus contenidos sino que es un "juicio de valor" emitido por un usuario particular para un documento específico.

La Web Cooperativa, al tener asociado cada usuario a un perfil, puede asignar a cada par `(perfil, documento)` un nivel de utilidad. El agente asignado a cada usuario sería el responsable de determinar dicho nivel de utilidad. Este proceso de evaluación, para ser verdaderamente práctico, debería determinarse de una forma implícita (únicamente "observando" el comportamiento del usuario, sin necesidad de interrogarle). Por otro lado, el nivel de utilidad no sería asignado al documento como un todo sino a pasajes individuales dentro de un mismo documento[1].

Ya se ha visto que la mayor parte de iniciativas relacionadas con la valoración de recursos por parte de los usuarios requieren una participación voluntaria con los problemas que esto conlleva. Sin embargo, también se han presentado algunas experiencias interesantes en el campo de la valoración implícita que han mostrado que es factible. La segunda opción es preferible de cara a una implementación práctica.

9.3.2 Recuperación de información para el maestro

Un agente de la Web Cooperativa tendría dos formas de obtener información para su maestro:

- Buscar información para satisfacer una consulta.
- Explorar en representación del usuario para recomendarle documentos desconocidos.

Para poder llevar a cabo ambas tareas se pretende emplear dos técnicas bien conocidas: Filtrado Colaborativo y Recomendación por Contenidos. En la Web Cooperativa, si el agente empleara filtrado colaborativo recomendaría al usuario documentos a los que usuarios de su mismo perfil han otorgado un elevado nivel de utilidad.

Por otro lado, si emplease recomendación por contenidos proporcionaría documentos relacionados conceptualmente con el perfil del usuario, con una consulta o con un documento de partida, independientemente del nivel de utilidad que pudieran tener asociado.

---

[1] J. Allan realizó un estudio que *"apoya claramente la hipótesis de que los documentos largos contienen información que diluye el feedback [la valoración del usuario]. Recortar estos documentos seleccionando un pasaje adecuado tiene un acentuado impacto en la eficiencia. (Allan 1995)."* En este caso no se reduciría un documento a un único pasaje sino que se extraería y trataría individualmente cada pasaje del texto.

Los agentes de la Web Cooperativa utilizarían un híbrido de ambas técnicas ya que esta forma de actuar facilita la localización de nuevos recursos en una comunidad incipiente (Burke 1999), aquella en la que aún no se han evaluado muchos documentos. En el siguiente punto se presentan ejemplos ilustrativos de ambos modos de funcionamiento del sistema.

9.4 Aplicaciones y limitaciones de la Web Cooperativa

En esta propuesta existen dos mecanismos de recuperación de información; el primero es comparable a los actuales motores de búsqueda mientras que el otro exploraría la Web en búsqueda de información que pudiera recomendar a los usuarios.

El primer sistema permitiría "consultas" similares a las descritas a continuación:

- "Encuentra documentos con el término `estrella`". Al tratarse de un término muy genérico el sistema no debería proporcionar ningún resultado sino indicar al usuario términos relacionados con el original en función del contexto. Así, podría ofrecer contextos que contuvieran, cada uno, conceptos como `Star Wars`, `astronomía`, `cine`, `música pop`, etc. Obviamente, un aspecto muy importante sería la interfaz que permitiría visualizar tales opciones.
- Encontrar documentos relacionados con una sentencia, párrafo o documento seleccionado por el usuario. El usuario introduciría un fragmento de texto o un *URI* y el sistema procedería a clasificar dicha información en una rama del árbol taxonómico retornando documentos de esa rama (o de ramas vecinas). De nuevo, en caso de que el texto de partida fuera excesivamente genérico no se proporcionarían resultados sino sugerencias para refinar la búsqueda.

Por supuesto, esto es sólo un primer esbozo del sistema de búsqueda; aspectos fundamentales para el mismo serían las técnicas de visualización de datos, así como aquellas para explorar los árboles taxonómicos u ordenar los resultados en función del usuario.

El sistema de recomendación funcionaría de forma ligeramente distinta; se trataría, básicamente, de un asistente personal que ayudaría al usuario realizando tareas como las siguientes:

- Buscar información en representación del usuario. El usuario proporcionaría al agente algunas consultas como las presentadas antes para que las procesara y extrajera un conjunto reducido de resultados.
- Recomendar documentos no solicitados pero interesantes. Para llevar a cabo esta tarea el asistente debería buscar documentos similares a otros procesados recientemente por el usuario así como intercambiar información con agentes similares; de esta forma sería posible satisfacer demandas latentes de información.

Si se compara la Web Cooperativa con la Web actual y con la Web Semántica está claro que esta propuesta proporciona menos resultados que los motores de búsqueda tradicionales aunque mucho más relevantes puesto que se están empleando taxonomías conceptuales. Por otro lado, al obtener semántica a partir del texto completo de los documentos la Web Cooperativa permite consultas difíciles para la Web Semántica a menos que se proporcione una ontología capaz de describir todos los conceptos y relaciones existentes, algo imposible en la mayor parte de los casos (p. ej., ¿Sería posible desarrollar una ontología lo suficientemente sutil como para describir la Informática y permitir cualquier consulta concebible?)

Por supuesto, consultas admisibles en la Web Semántica como "*Encontrar el artículo más reciente sobre SHOE en el que James Hendler figure como coautor* (Denker *et al* 2001)" no podrían ser resueltas satisfactoriamente en la Web Cooperativa. Por esa razón la Web Cooperativa se propone como complemento de la Web Semántica y no como sustituto.

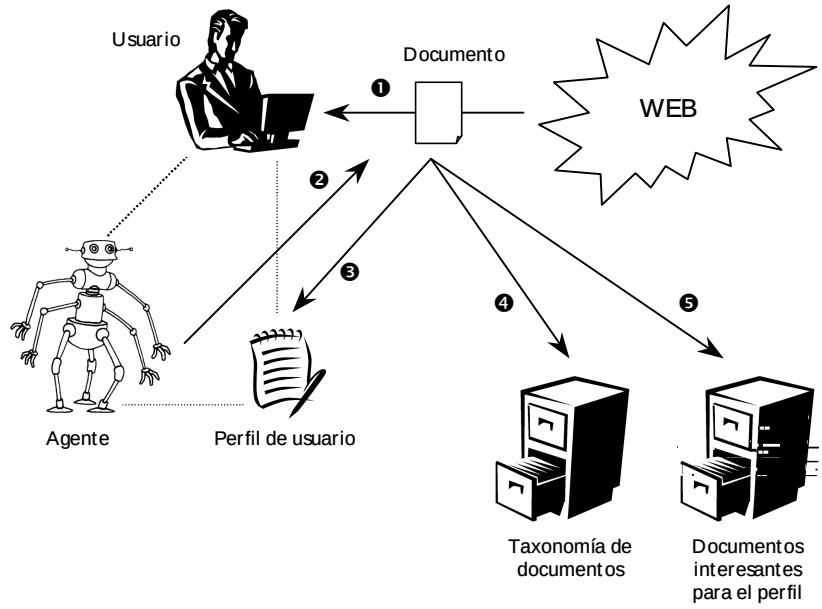


Fig. 8 Funcionamiento básico de la Web Cooperativa.

El usuario navega por la Web de la manera usual y descarga un documento ❶, su agente observa todas sus acciones y en función de éstas valora el interés del documento para el usuario ❷. Una vez evaluada la relevancia del documento, el agente actualiza el perfil del usuario sobre la base de la nueva información ❸, clasifica el documento en caso necesario dentro de la taxonomía de documentos ❹ (que estaría alojada en un servidor central) y agrega el documento, en caso de que la valoración sea positiva, a un "repositorio" de documentos de interés para el perfil del usuario que representa ❺ (también alojado en un servidor central).

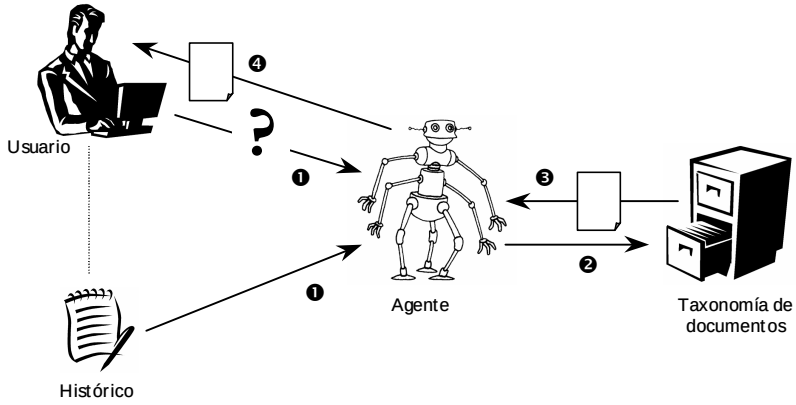


Fig. 9 Resolución de consultas y recomendación por contenidos en la Web Cooperativa.

Los agentes de la Web Cooperativa pueden resolver consultas de los usuarios además de explorar en representación de los mismos (recomendar documentos cuyos contenidos pueden ser interesantes). El agente puede examinar el histórico de navegación del usuario o recibir una consulta ❶. Con esta información el agente lleva a cabo una exploración taxonómica ❷, es decir, clasifica dentro de la taxonomía conceptual los datos de partida y obtiene como resultados documentos próximos en el dendrograma ❸. Estos documentos son proporcionados al usuario como recomendaciones en caso de que el agente haya actuado *motu proprio* o como resultados de una consulta ❹.

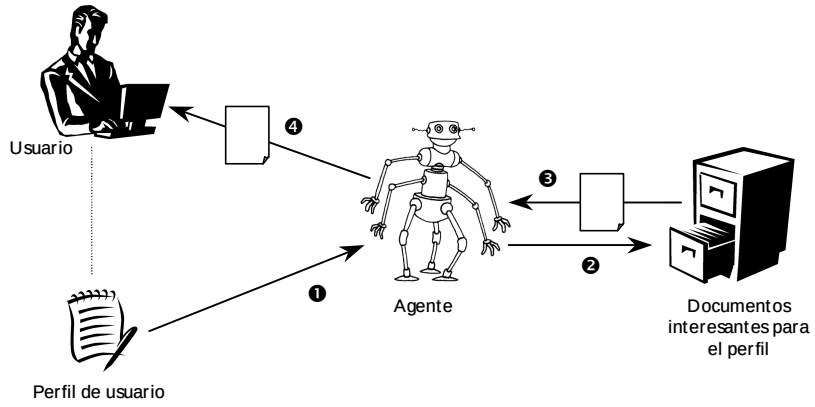


Fig. 10 Recomendación por filtrado colaborativo en la Web Cooperativa.

Los agentes de la Web Cooperativa pueden recomendar documentos de interés para el usuario basándose en las preferencias de usuarios similares. Periódicamente, cada agente accedería a distintos repositorios ❷ en función del perfil de su maestro ❶. De cada repositorio se obtendrían una serie de documentos potencialmente interesantes ❸ que serían presentados al usuario como recomendaciones ❹.

10 ¿Qué NO es la Web Cooperativa?

A la luz de lo visto hasta ahora es posible proporcionar una definición para la Web Cooperativa:

> *"La Web Cooperativa es una capa situada directamente sobre la Web actual con el fin de dotarla de semántica de manera global, automática, transparente e independiente del idioma. Requiere la participación de los usuarios pero no de forma consciente y directa sino indirectamente a través de agentes autónomos y cooperantes. La Web Cooperativa se apoya sobre el uso de conceptos y taxonomías documentales; unos y otras pueden obtenerse, sin intervención humana, a partir del texto libre de los documentos."*

En los apartados anteriores se ha planteado el problema, se ha situado en un contexto más amplio y se han mostrado iniciativas que han tratado de resolverlo parcialmente y la forma en que éstas han inspirado y motivado al autor en la propuesta de la Web Cooperativa dentro de la cual, como se verá más adelante, se enmarca su tesis. Existen, sin embargo, distintos proyectos que no estando relacionados con esta propuesta podrían parecer, engañosamente, similares; la finalidad de este apartado es diferenciar la propuesta de Web Cooperativa de estas otras.

10.1 La Web Cooperativa NO es la Web Semántica

La Web Cooperativa pretende extraer semántica de los documentos existentes en la Web, "clasificar" los documentos en una taxonomía o dendrograma y utilizar agentes. A la vista de esto es posible intentar compararla con la Web Semántica; sin embargo, eso sería un error puesto que las diferencias entre ambas iniciativas son enormes.

La Web Semántica requiere ontologías, sean estas construidas automáticamente o desarrolladas por un ser humano; dichas ontologías definen clases y relaciones que permiten etiquetar documentos para, así, facilitar un proceso de inferencia a los agentes de la Web Semántica.

De este modo, en la Web Semántica hasta que un concepto no está recogido en una ontología no existe pues no puede ser nombrado de ningún modo. Por otro lado, ya se ha

comentado anteriormente que la Web Semántica, a pesar de su nombre, ofrece a la Web más metasemántica que semántica.

La Web Cooperativa, por otro lado, no emplea ontologías, sólo conceptos. Este enfoque es mucho más simple puesto que no interesa explicitar en modo alguno las relaciones entre los conceptos. Esto no quiere decir que la Web Cooperativa ignore las relaciones entre conceptos sino que son manipuladas implícitamente.

Como ya se dijo, cada pasaje de cada documento es una secuencia de conceptos y el autor cree que dichas secuencias conceptuales pueden ser procesadas de modo similar a como el ADN es procesado para establecer clasificaciones de seres vivos. Esta clasificación conceptual automática, en caso de ser posible, sería capaz de separar documentos de un modo similar a como haría un ser humano dejando patentes las relaciones implícitas entre conceptos.

Por otro lado, los agentes de la Web Semántica y la Web Cooperativa tendrían misiones muy distintas. Los primeros tendrían como finalidad procesar documentos etiquetados "semánticamente" y realizar inferencias. Los segundos procesarían documentos no etiquetados, aprenderían de sus maestros e intercambiarían información entre ellos con el objetivo de recomendar información interesante.

Por todo ello, aun cuando tanto la Web Semántica como la Web Cooperativa emplean agentes, elementos semánticos y establecen algún tipo de catalogación de documentos, se trata de propuestas totalmente distintas (aunque como se ha señalado anteriormente complementarias).

10.2 La Web Cooperativa NO son las categorías *dmoz* o *Yahoo!*

Un aspecto vital de la Web Cooperativa es la clasificación de documentos en taxonomías o dendrogramas. Tales dendrogramas permitirían mostrar las relaciones conceptuales existentes entre los documentos de forma análoga a como se visualizan las relaciones que hay entre distintas especies biológicas y deberían obtener, de forma automática, "categorías" de documentos muy similares a las que podría establecer un ser humano.

Estas categorías pueden recordar a las disponibles en directorios como *dmoz*[1], *looksmart*[2] o *Yahoo!*[3]; no obstante, aun cuando es posible un parecido superficial, las diferencias de fondo entre las taxonomías documentales de la Web Cooperativa y las de estos directorios son notables.

Recuérdese que la propuesta que se plantea en este trabajo pretende generar de forma totalmente automática, no supervisada e independiente del idioma una o más taxonomías para los documentos disponibles en la Web; la estrategia seguida por los directorios es, sin embargo, muy distinta.

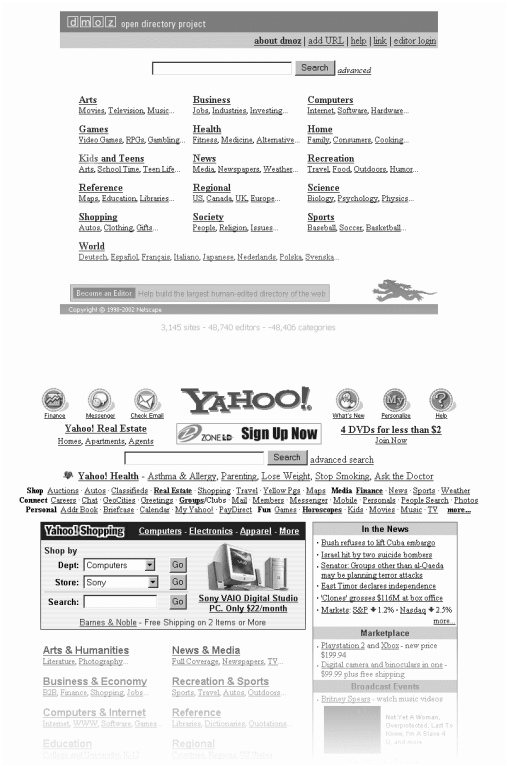
Todos los directorios requieren supervisión humana tanto para la creación de las categorías como para la asignación de documentos a las mismas. La forma en que se llevan a cabo estas tareas varía de un directorio a otro pero en ningún caso pueden realizarse de forma totalmente automática.

---

[1] `http://www.dmoz.org`

[2] `http://www.looksmart.com`

[3] `http://www.yahoo.com`



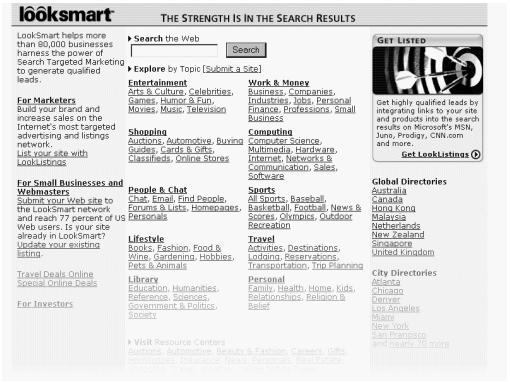


Fig. 11 Directorios *dmoz*, *looksmart* y *Yahoo!*

Página principal de los tres principales directorios de la Web a fecha de 20 de mayo 2002. Los tres comparten básicamente la misma estructura de categorías y proporcionan un motor de búsqueda que puede operar, al menos, sobre los documentos indexados en el directorio.

Yahoo! y *looksmart* cuentan con una plantilla dedicada a construir sus respectivos directorios. Estos empleados reciben sugerencias de los usuarios de la Web para añadir sitios web al directorio (no así para crear nuevas categorías). Esta estrategia está muy limitada y difícilmente puede desarrollar un directorio que abarque la totalidad, o cuando menos una parte importante, de la Web y que sea, al mismo tiempo, de calidad.

Además, esta forma de construir directorios tiende a "prostituirse" al promocionar determinadas categorías o documentos dentro de una categoría a cambio de una suma de dinero. Este hecho, aunque comercialmente justificable, sin duda degrada la utilidad que el directorio pudiera tener para los usuarios lo cual hace este método desaconsejable.

En el otro extremo se sitúan iniciativas como la de *dmoz*, también conocido como *Open Directory Project* u *ODP*. Se trata de un directorio desarrollado por una comunidad de usuarios que actúan desinteresadamente, de un modo similar a como se desarrolla el Software Libre. Cada categoría está gestionada por uno o más editores que revisan las sugerencias enviadas por los usuarios (documentos o propuestas de nuevas categorías), proporcionan una descripción para las mismas y organizan los documentos que aparecen en la categoría creando, en caso necesario, subcategorías.

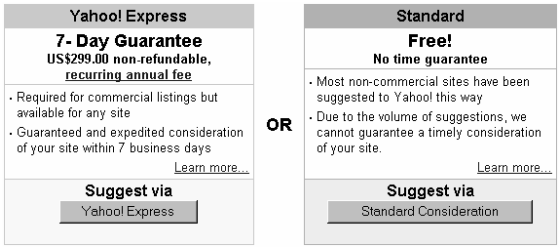


Fig. 12 Sugerencia de un nuevo sitio web para el directorio Yahoo!

Yahoo! permite a los usuarios sugerir nuevas entradas al directorio. A cambio de 299 dólares se "tomará en consideración" la "sugerencia" en sólo 7 días. La segunda opción (gratuita) no garantiza la inclusión del enlace en el directorio en ningún momento.

Esta estrategia colaborativa y altruista es superior a la empleada por directorios comerciales como los anteriormente mencionados puesto que es más fácilmente escalable y menos susceptible a la "corrupción". Sin embargo, a pesar de su mayor escalabilidad sigue sin poder abarcar una parte importante de la Web[1].

Por tanto, aun cuando las taxonomías que se proponen para la Web Cooperativa podrían coincidir en muchas ocasiones con las categorías disponibles en directorios como *dmoz* o *Yahoo!*, existen diferencias muy claras entre ambas iniciativas: los directorios son supervisados por humanos mientras que las taxonomías de la Web Cooperativa serían obtenidas de forma totalmente automática.

10.3 La Web Cooperativa NO es la Web Colaborativa

El término elegido para la propuesta, Web Cooperativa, tal vez no haya sido excesivamente afortunado puesto que puede llevar a confusión con algunas iniciativas calificadas, en ocasiones, como Web Colaborativa.

A diferencia de la Web Semántica que da nombre a una serie de líneas de investigación bien delimitadas, se ha empleado el término "Web Colaborativa" en varios proyectos, académicos y comerciales, que tienen poco o nada que ver entre sí ni con la propuesta de Web Cooperativa. A continuación se citan algunas de las aplicaciones calificadas en una u otra ocasión como Web Colaborativa.

GroupWeb (Greenberg y Roseman 1996) introduce el concepto de navegación colaborativa (*collaborative web browsing*) al presentar un sistema que permite a varios usuarios navegar de forma conjunta (recomendarse enlaces, seguir la ruta de navegación de otro usuario, explorar de forma combinada, etc.) Posteriormente, surgieron otra serie de iniciativas muy similares. Todos estos proyectos son, sin embargo, aplicaciones para trabajo en grupo y no sistemas de recuperación de información.

Sparrow Web (Chang 1998) fue un proyecto desarrollado en *Xerox PARC* que permitía a varios usuarios modificar una página web directamente mediante su navegador. Esta iniciativa se parece bastante a la idea de los *Wikis*[2] y, como se puede ver, no tiene ningún punto en común con la propuesta aquí descrita.

Kovács y Micsik (2000) emplean el término "Web Colaborativa" para hacer referencia a aplicaciones web que permiten el trabajo simultáneo de varios individuos. Sin embargo, describen aplicaciones relativamente tradicionales de filtrado colaborativo en *USENET*, Web y en una biblioteca digital (empleando en todos los casos valoración explícita) así como un sistema de encuestas y votaciones.

En resumen, la Web Colaborativa permite la colaboración de individuos de manera transparente ya sea para modificar documentos, explorar la Web o intercambiar información. Sin embargo, en el caso de la Web Cooperativa las entidades que cooperan son agentes que actúan en representación de los usuarios. Esta cooperación resulta transparente para el usuario que sigue empleando la Web de la manera usual.

---

[1] El directorio *dmoz* tenía indexados 3.429.012 ($3{,}4 \cdot 10^6$) sitios web a fecha de 28 de mayo de 2002, contando para ello con 49.030 editores; *Google* tenía indexadas 2.073.418.204 páginas ($2{,}1 \cdot 10^9$). Teniendo en cuenta que un directorio almacena una única página por sitio, *dmoz* está aún a tres órdenes de magnitud del volumen de documentos procesados por un sistema automático como *Google*.

[2] Un *Wiki* es un sitio web donde las páginas pueden ser editadas por cualquier visitante. Cualquier usuario puede ayudar a mejorar el sitio o plantear sus dudas, editando la página web, esperando que otro usuario las resuelva.

11 Formulación definitiva del problema y de la tesis

Parece innecesario decir que aún no existe ninguna implementación de la Web Cooperativa; sin embargo, utilizándola como una "vista desde 20.000 pies" resulta muy útil al plantear toda una serie de problemas interesantes:

- ¿Qué acciones del usuario sobre un documento son altamente discriminantes para determinar implícitamente su relevancia?
- ¿Es posible utilizar tales reglas de un modo eficiente dentro de un navegador web para determinar la relevancia del documento visualizado en un momento dado?
- ¿Es posible clasificar textos libres empleando métodos tomados de la biología computacional?
- ¿Es posible obtener un "pseudo-ADN" a partir de texto escrito en un lenguaje natural?
- Si existiera ese pseudo-ADN, ¿sería posible combinarlo, mutarlo o construir "algo" a partir del mismo como sucede con el ADN real?
- ¿Se debe suponer que idiomas distintos constituyen "bioquímicas" diferentes?
- ¿Cómo se daría el salto desde ese pseudo-ADN a los conceptos?
- ¿En qué forma podría un agente realizar búsquedas eficientes sobre una taxonomía de documentos construida a partir de ese pseudo-ADN?
- ¿Cómo obtendrían, representarían y almacenarían los agentes el perfil de sus maestros?
- ¿Cómo y dónde se comunicarían los agentes entre sí?
- ¿Qué información sobre el perfil de los respectivos maestros podría ser intercambiada?

Así pues, la Web Cooperativa involucra muy diversos aspectos: tratamiento de lenguaje natural, evaluación implícita de documentos, agentes *software*, interacción persona-ordenador, usabilidad o privacidad. En este trabajo se prescinde de lo que serían las "capas superiores" de la Web Cooperativa y se delimita aún más el problema encuadrándolo dentro del campo del procesamiento de lenguaje natural por medios estadísticos.

De este modo, a partir del problema original se formula otro más concreto sobre el que finalmente versa el presente trabajo:

> *La cantidad de texto no estructurado disponible en la Web seguirá aumentando y, a pesar de sus inconvenientes, el método preferido por la mayor parte de usuarios para recuperar información continuarán siendo las consultas formuladas en lenguajes naturales. En ambos casos (publicación y consulta) será inevitable un uso generalmente ambiguo de los distintos idiomas y la presencia de errores tipográficos, ortográficos o gramaticales.*

En relación con dicho problema, el autor sostiene la siguiente tesis:

> *Se puede obtener para los distintos n-gramas, g_i, de un texto escrito en cualquier idioma una medida de su significatividad, s_i, distinta de la frecuencia relativa de aparición de los mismos en el texto, f_i, pero calculable a partir de la misma. Esta métrica de la significatividad intradocumental de los n-gramas permite asociar a cada documento, d_i, un único vector, v_i, susceptible de comparación con cualquier otro vector obtenido del mismo modo aun cuando sus respectivas longitudes puedan diferir. Puesto que tales vectores almacenan ciertos aspectos de la semántica subyacente a los textos originales, el mayor o*

menor grado de similitud entre los mismos constituye un indicador de su nivel de relación conceptual, facilitando la clasificación y categorización de documentos, así como la recuperación de información. Asimismo, cada vector individual es capaz de transformar el texto original a partir del cual fue obtenido dando lugar a secuencias de palabras clave y resúmenes automáticos.

Siendo ésta su versión resumida:

Una única técnica sencilla, basada en el uso de vectores de n-gramas de longitud variable, independiente del idioma y aplicable a diversas tareas de tratamiento de lenguaje natural con resultados similares a los de otros métodos 'ad hoc' es viable.

A lo largo de los siguientes capítulos se procederá a describir los fundamentos de la nueva técnica propuesta por el autor y su relación con otras existentes. Se demostrará que, efectivamente, es posible obtener de forma sencilla y automática representaciones de documentos que conservan aspectos semánticos de los mismos con independencia del idioma. Se describirán las aplicaciones de la técnica a diversas tareas de procesamiento de lenguaje natural. Y, para finalizar, se expondrán las conclusiones a las que ha llegado el autor del trabajo y se esbozarán posibles líneas de trabajo futuro.

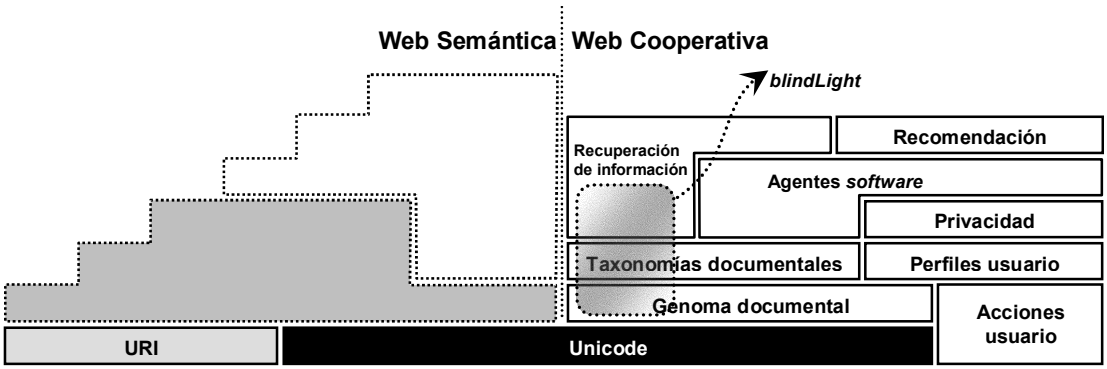


Fig. 13 Web Semántica vs. Web Cooperativa y relación de esta última con *blindLight*.

CAPÍTULO 2

TÉCNICAS ESTADÍSTICAS PARA PROCESAMIENTO DE LENGUAJE NATURAL

La cantidad de texto electrónico disponible en distintos idiomas es enorme. Sin embargo, dichos textos son normalmente "planos" y en muchas ocasiones pueden contener errores tipográficos, ortográficos o gramaticales. A pesar de ello encierran una cantidad enorme de información que los usuarios podrían explotar una vez "tamizada". Para ello es necesario disponer de técnicas de procesamiento de lenguaje natural que permitan la clasificación, categorización, extracción y recuperación de información. Dichas técnicas deberían ser sencillas, robustas y aplicables a múltiples idiomas sin recurrir a conocimiento lingüístico alguno. A lo largo de este capítulo se describirán varias de estas técnicas poniendo énfasis en métodos puramente estadísticos. Así, se describirá el modelo vectorial estudiando su aplicación a la clasificación y categorización de documentos así como a la recuperación de información. Posteriormente se analizará el uso de n-gramas de caracteres en dicho modelo y se continuará con las técnicas Acquaintance y Highlights con las que la propuesta del autor muestra ciertas similitudes.

1 Sobrecarga de información y Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (PLN) es el conjunto de técnicas algorítmicas que tienen como objeto la manipulación y generación de muestras de lenguaje humano tanto en su manifestación escrita como oral. Ejemplos de técnicas de PLN son la generación de habla a partir de texto, el reconocimiento del habla, la traducción automática o la recuperación de información.

En el capítulo anterior se limitó el problema objeto de estudio al procesamiento de texto natural en condiciones "extremas" (grandes volúmenes de texto, múltiples idiomas, ambigüedad, errores tipográficos, ortográficos y gramaticales) por lo que, en este trabajo, no resulta de interés ninguna de las técnicas PLN relacionadas con el habla.

Por otro lado, es posible desglosar el problema de partida, la "sobrecarga de información", en una serie de tareas bien definidas: **categorización** (asignación de un

documento a una categoría previamente conocida), **clasificación** (agrupación de documentos con características similares), **recuperación de información** (localización, dentro de una colección de documentos, de un subconjunto relevante para una consulta formulada por un usuario) y **destilación de información** (extracción de palabras clave, obtención de resúmenes automáticos, respuesta de preguntas, etc.) Así, tampoco se tratarán en este trabajo técnicas PLN que tengan como objeto la creación[1] de muestras de lenguaje.

Aun así existe un amplio repertorio de métodos para resolver cada una de las tareas anteriores y es posible implementar herramientas que utilicen varias de dichas técnicas para afrontar el problema genérico de procesamiento de texto no estructurado en la Web. No obstante, debido a las especiales condiciones en las que dicho procesamiento debería llevarse a cabo el autor considera que las técnicas a emplear deberían verificar los siguientes requisitos:

- Independencia del idioma (aplicables a diversos lenguajes humanos sin cambios o con cambios mínimos).
- Utilización únicamente de métodos estadísticos simples (conocimiento "cero").
- Alta tolerancia al "ruido" (capacidad para trabajar sobre documentos "contaminados[2]" o con errores de cualquier tipo).
- Escalabilidad (posibilidad de ser aplicadas sobre colecciones de documentos muy grandes y de crecimiento continuo).

Por ello no se estudirá ninguna técnica que requiera el uso de "artefactos" lingüísticos como *stemmers*, etiquetado *POS*[3] o desambiguación puesto que éstos requieren conocimientos del idioma en que están escritos los documentos. La razón para esta limitación del abanico de técnicas aceptables es simple: siempre es más sencillo conseguir muestras de texto plano para cualquier lengua que el correspondiente conocimiento lingüístico sobre la misma. Es más, el autor considera indispensable disponer de métodos simples y robustos aplicables en semejantes condiciones como paso previo al desarrollo de técnicas PLN más elaboradas.

A lo largo de este capítulo se analizarán muy brevemente las características de métodos aplicables a cada una de las tareas descritas y que verifican los cuatro primeros requisitos. Sin embargo, a todo lo anterior habría que añadir un quinto requisito, a saber, el uso de una única técnica para afrontar cada una de las tareas anteriores. Recuérdese que el autor sostiene en su tesis no sólo la posibilidad de desarrollar semejante técnica sino que ésta ofrecerá para cada una de las tareas anteriores resultados comparables a los obtenidos

---

[1] Entendiendo la creación como la producción "desde cero" de textos escritos en un lenguaje natural.

[2] Artículos *USENET* a los que no se han eliminado las cabeceras o documentos *HTML* a los que no se ha podido limpiar todo el código *Javascript* son ejemplos de documentos "contaminados".

[3] *Part-Of-Speech* (*POS*) es el papel que una palabra juega en una producción (por ejemplo, sustantivo, verbo, adjetivo, etc.) Un etiquetador *POS* recibe una producción en un lenguaje natural y produce una salida en la que cada palabra está "etiquetada" con uno o más tipos. Por ejemplo, la oración **Él te vino a ver** podría etiquetarse de la siguiente manera: **Él** [Pronombre personal] **te** [Pronombre personal] **vino** [Verbo principal indicativo] **a** [Preposición] **ver** [Verbo principal infinitivo]. En cambio, **El té y el vino están pasados** sería etiquetada de la forma siguiente: **El** [Artículo definido] **té** [Nombre común] **y** [Conjunción coordinada] **el** [Artículo definido] **vino** [Nombre común] **están** [Verbo principal indicativo] **pasados** [Verbo principal participio]. Este tipo de información resulta enormemente útil pero no se puede obtener de manera automática sin conocimiento del lenguaje en cuestión.

con técnicas específicas. Dicha técnica, denominada *blindLight*, se describirá en el siguiente capítulo.

2 El modelo vectorial de documentos

A excepción de la obtención de resúmenes automáticos[1], todas las técnicas PLN que resultan de interés para el problema que nos ocupa y con las que tiene relación la nueva técnica propuesta por el autor tienen dos puntos en común: (1) Requieren una definición de asociación (o similitud) entre dos documentos y (2) precisan una forma de representación de los documentos que permita el cálculo de dicha medida de asociación.

Así, en el caso de la clasificación se desea construir subconjuntos de documentos que exhiban unas características comunes aunque diferentes de las del resto de grupos. Dicho de otro modo, deben encontrarse grupos que maximicen la similitud intragrupal al tiempo que minimicen la similitud intergrupal. La categorización, por su parte, se reduce a determinar qué categoría (que se representará de un modo análogo a los documentos) se encuentra más próxima al documento a categorizar. Por último, los sistemas de recuperación de información reciben consultas (esto es, documentos extremadamente cortos producidos por el usuario) y retornan aquellos documentos de la colección que se encuentran más próximos a las mismas.

Así pues, en lugar de analizar varias técnicas PLN de forma aislada se va a tratar de ofrecer una visión global de las mismas. Para ello se estudiarán las distintas formas en que se puede representar un documento así como las posibles medidas de asociación con cada tipo de representación.

La forma más sencilla de representar un documento es mediante un conjunto de palabras. Aquellas que aparecen en el documento pertenecerán al conjunto y las que no se utilizan, obviamente, no. En este modelo, denominado booleano, las consultas no se representan del mismo modo que los documentos sino bajo la forma de expresiones lógicas que combinan palabras (que presumiblemente se utilizan en los documentos) y los operadores `AND`, `OR` y `NOT`.

```
(information AND retrieval) OR ir
```

Fig. 14 Ejemplo de consulta para un modelo booleano.

El modelo booleano es muy simple, de hecho demasiado: todas las palabras de un documento son consideradas igualmente importantes y las consultas retornan o bien demasiados documentos o bien muy pocos. Por otro lado, puesto que no existe el concepto de similitud no es posible determinar qué documentos satisfacen mejor la consulta, es decir, el funcionamiento es dicotómico: hay documentos que no satisfacen la consulta y otros que sí (aquellos que aparecen en la lista de resultados).

Sin embargo, es posible definir medidas de asociación entre este tipo de representaciones de documentos y, por tanto, similitudes. La más simple de tales medidas es la siguiente:

$$\left|X \cap Y\right| \qquad (1)$$

[1] En realidad, algunas técnicas de extracción de resúmenes se han construido sobre sistemas de recuperación de información y, por tanto, también comparten ambas características.

Que no es más que el número de palabras que aparecen tanto en el documento *X* como en el documento *Y*. Sin embargo, esta medida no tiene en cuenta el número de términos de cada documento y puede resultar engañosa al comparar resultados obtenidos para parejas de documentos con longitudes muy diferentes. Los siguientes coeficientes se basan en el anterior pero incluyen más información sobre los dos documentos a comparar:

$$2\frac{|X \cap Y|}{|X|+|Y|} \quad (2)$$

Coeficiente de Dice

$$\frac{|X \cap Y|}{|X \cup Y|} \quad (3)$$

Coeficiente de Jaccard

$$\frac{|X \cap Y|}{|X| \times |Y|} \quad (4)$$

Coseno

$$\frac{|X \cap Y|}{\min(|X|,|Y|)} \quad (5)$$

Coeficiente de solapamiento

La utilización de tales coeficientes permite mejorar el modelo en varios aspectos. En primer lugar, las consultas pueden representarse como conjuntos de palabras exactamente igual que los documentos. En segundo lugar, es posible obtener un valor numérico que indique cuán fuerte (o débil) es la relación entre dos documentos (y por tanto entre un documento y una consulta).

A pesar de estas ventajas, la representación de los documentos como vectores booleanos no es totalmente satisfactoria puesto que, como ya se dijo, todas las palabras resultan igualmente relevantes lo cual no es realista. Es mucho más conveniente asignar a cada palabra un valor real, un "peso", que indique la importancia de la misma dentro de dicho documento.

Este nuevo modelo, conocido como **modelo vectorial** (Salton y Lesk 1965), considera cada documento de una colección como un vector de pesos en un espacio de *T* dimensiones donde *T* es el número de términos distintos que aparecen en la colección.

$$D_i = (d_{i1}, d_{i2}, d_{i3}, \ldots, d_{it})$$

Fig. 15 Un documento en un espacio vectorial de T dimensiones.

d_{ij} es el peso del término *j-simo* para el documento D_i.

Para calcular el peso de un término en un documento existen distintas alternativas pero en todos los casos se tiene en cuenta lo siguiente:

- La frecuencia de aparición del término en el propio documento, *tf* (Luhn 1957). Los términos que más se repiten en un documento son, en principio, más relevantes que los que se emplean menos.

- El número de documentos de la colección en los que aparece el término, *idf* (Karen Spärck-Jones 1972). Los términos más frecuentes en la colección serán menos relevantes que los más raros.
- La longitud del documento, a fin de garantizar que todos los documentos se comportan de modo similar con independencia de su longitud. En otras palabras, no hay relación entre la relevancia de un documento para una consulta y su longitud.

D1: Microsoft vs Google heats up
D2: Microsoft previews MSN Virtual Earth
D3: MSN to offer virtual Earth map service
D4: MSN joins Google in melding satellite imagery with search
D5: MSN Virtual Earth to take on Google Earth

Q: Google Earth

Fig. 16 Una colección de 5 documentos y una consulta.

Término	IDF	w_{D1i}	w_{D2i}	w_{D3i}	w_{D4i}	w_{D5i}	w_{Qi}
earth	0,33	0	0,33	0,33	0	0,66	0,33
google	0,33	0,33	0	0	0,33	0,33	0,33
heats	1	1	0	0	0	0	0
imagery	1	0	0	0	1	0	0
in	1	0	0	0	1	0	0
joins	1	0	0	0	1	0	0
map	1	0	0	1	0	0	0
melding	1	0	0	0	1	0	0
microsoft	0,50	0,50	0,50	0	0	0	0
msn	0,25	0	0,25	0,25	0,25	0,25	0
offer	1	0	0	1	0	0	0
on	1	0	0	0	0	1	0
previews	1	0	1	0	0	0	0
satellite	1	0	0	0	1	0	0
search	1	0	0	0	1	0	0
service	1	0	0	1	0	0	0
take	1	0	0	0	0	0,50	0
to	0,50	0	0	0,50	0	0,50	0
up	1	1	0	0	0	0	0
virtual	0,33	0	0,33	0,33	0	0,33	0
vs	1	1	0	0	0	0	0
with	1	0	0	0	1	0	0

Fig. 17 La colección anterior y la consulta representadas en un espacio vectorial.

El valor *idf* se ha simplificado como el inverso del número de documentos en que aparece cada término. El peso de cada término para cada documento es el producto de dicho *idf* por la frecuencia de aparición del término en el documento, así, el término `Earth` tiene un peso de 0.33 en todos los documentos a excepción del quinto pues en éste aparece dos veces y, en consecuencia, el peso es 0.66.

No parece necesario entrar en mayores detalles acerca del cálculo de los pesos para entender el funcionamiento del modelo: para cada documento de una colección se genera un vector de valores reales y en el caso de los sistemas de recuperación de información basados en el modelo vectorial se procede del mismo modo para las consultas recibidas por el sistema.

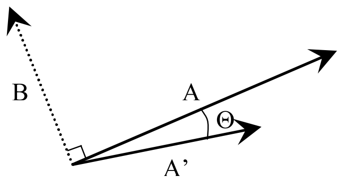
Así pues, dados dos vectores de pesos es posible, de manera análoga a como se hacía con vectores booleanos, obtener una medida numérica de su asociación. Para ello, pueden adaptarse algunas de la medidas de asociación mostradas antes, siendo una de las más populares la denominada **"función del coseno"** (véase Fig. 18).

$$\frac{\sum_{i=1}^{n} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}}$$

Fig. 18 Ecuación de la función del coseno entre los documentos *Q* y *D*.

q_i y d_i son el componente i-ésismo de los vectores *Q* y *D*, respectivamente. *n* es el número de términos distintos en la colección.

Esta medida admite una interpretación geométrica (véase Fig. 19) puesto que su valor numérico puede considerarse como el coseno del ángulo formado por los vectores de los documentos comparados. Así, un valor 0 implica que los vectores son ortogonales (esto es, no son similares) mientras que un valor 1 significa que los vectores forman un ángulo de 0º (es decir, son iguales o, más bien, muy parecidos).



d (A, A') ≅ 1 ➔ Θ ≅ 0º
d (A, B) = 0 ➔ Θ = 90º

Fig. 19 Interpretación geométrica de la función del coseno en un espacio bidimensional.

Los documentos A y A' tienen una similitud próxima a 1, es decir, forman un ángulo cercano a 0º y, por tanto, "apuntan" en la misma dirección. En cambio, la similitud entre A y B es 0, es decir, son ortogonales.

Término	w_{D1i}^2	w_{D2i}^2	w_{D3i}^2	w_{D4i}^2	w_{D5i}^2	w_{Qi}^2	$w_{D1i} \cdot w_{Qi}$	$w_{D2i} \cdot w_{Qi}$	$w_{D3i} \cdot w_{Qi}$	$w_{D4i} \cdot w_{Qi}$	$w_{D5i} \cdot w_{Qi}$
earth	0	0,11	0,11	0	0,44	0,11	0	0,11	0,11	0	0,22
google	0,11	0	0	0,11	0,11	0,11	0,11	0	0	0,11	0,11
heats	1	0	0	0	0	0	0	0	0	0	0
imagery	0	0	0	1	0	0	0	0	0	0	0
in	0	0	0	1	0	0	0	0	0	0	0
joins	0	0	0	1	0	0	0	0	0	0	0
map	0	0	1	0	0	0	0	0	0	0	0
melding	0	0	0	1	0	0	0	0	0	0	0
microsoft	0,25	0,25	0	0	0	0	0	0	0	0	0
msn	0	0,06	0,06	0,06	0,06	0	0	0	0	0	0
offer	0	0	1	0	0	0	0	0	0	0	0
on	0	0	0	0	1	0	0	0	0	0	0
previews	0	1	0	0	0	0	0	0	0	0	0
satellite	0	0	0	1	0	0	0	0	0	0	0
search	0	0	0	1	0	0	0	0	0	0	0
service	0	0	1	0	0	0	0	0	0	0	0
take	0	0	0	0	0,25	0	0	0	0	0	0
to	0	0	0,25	0	0,25	0	0	0	0	0	0
up	1	0	0	0	0	0	0	0	0	0	0
virtual	0	0,11	0,11	0	0,11	0	0	0	0	0	0
vs	1	0	0	0	0	0	0	0	0	0	0
with	0	0	0	1	0	0	0	0	0	0	0

|D1|	|D2|	|D3|	|D4|	|D5|	|Q|
1,83	1,24	1,88	2,68	1,49	0,47

D1·Q	D2·Q	D3·Q	D4·Q	D5·Q
0,11	0,11	0,11	0,11	0,33

|D1|·|Q|	|D2|·|Q|	|D3|·|Q|	|D4|·|Q|	|D5|·|Q|
0,86	0,58	0,88	1,26	0,70

D1, Q	D2, Q	D3, Q	D4, Q	D5, Q
0,13	0,19	0,13	0,09	0,47

Fig. 20 Cálculo de la función del coseno entre la consulta y los documentos anteriores.

? Google Earth

(0,47) **D5:** MSN Virtual Earth to take on Google Earth
(0,19) **D2:** Microsoft previews MSN Virtual Earth
(0,13) **D1:** Microsoft vs Google heats up
(0,13) **D3:** MSN to offer virtual Earth map service
(0,09) **D4:** MSN joins Google in melding satellite imagery with search

Fig. 21 Lista ordenada de resultados al realizar la consulta sobre la colección.

Otra ventaja de la función del coseno radica en el hecho de que no es necesario normalizar los pesos de los términos en función de la longitud de los documentos puesto que lo que se "mide" es el ángulo formado por los vectores (véase Fig. 19).

Mediante la utilización de vectores de pesos para representar documentos (y consultas) y el uso de la función del coseno u otra similar es sencillo implementar sistemas de recuperación de información (véanse Fig. 16, Fig. 17, Fig. 20 y Fig. 21). Representando de manera vectorial las características comunes de un conjunto de documentos se puede utilizar este modelo también para realizar categorización y clasificación. En el modelo booleano resulta muy sencillo encontrar un nuevo vector con las características comunes de dos o más documentos, basta con realizar la intersección de todos ellos para obtener un nuevo vector que contenga aquellos términos comunes a todos ellos. En el caso del modelo vectorial la solución es similar: se necesita calcular el **centroide** del conjunto de vectores, es decir, obtener un nuevo vector de pesos donde cada peso será la media de los pesos de los distintos vectores del conjunto.

En definitiva el modelo vectorial ofrece un modo de representar documentos y consultas mediante vectores de pesos, una serie de medidas para determinar la asociación entre dichos vectores y un modo de calcular un nuevo vector (centroide) para representar características comunes de un conjunto de vectores (o lo que es lo mismo, un grupo de documentos). Todo ello permite aplicar el modelo vectorial a tres de las cuatro tareas antes mencionadas[1]:

- **Categorización:** un sistema basado en el modelo vectorial puede ser "entrenado" de una forma muy sencilla. Dado un conjunto de documentos de entrenamiento etiquetados se calcula el centroide del conjunto de documentos pertenecientes a cada categoría. Una vez hecho esto, la categorización resulta tan sencilla como determinar qué centroide (categoría) es el más próximo al vector del documento a categorizar.
- **Clasificación:** dado el conjunto de vectores para los documentos a clasificar y una medida de asociación es posible implementar cualquiera de los algoritmos clásicos de clasificación (*clustering* aglomerativo, partición, *k*-vecinos, etc.)
- **Recuperación de información:** se obtiene un conjunto de vectores para todos los documentos de la colección, al recibir una consulta se genera un vector para la consulta y se calcula la similitud entre éste y los vectores de la colección proporcionando como resultado una lista de documentos ordenada inversamente por su similitud a la consulta.

[1] Y también a la extracción de resúmenes: (1) cada sentencia de un documento se representa mediante un vector, (2) se calcula el centroide de estos vectores, (3) se calcula la distancia coseno de cada vector sentencia al centroide y (4) se extraen las sentencias más próximas al centroide como resumen del texto. No obstante, esta técnica, aunque sencilla, no es la mejor posible para obtener resúmenes extractivos.

Además, estas implementaciones cumplen tres de los cuatro requisitos deseables para resolver el problema del tratamiento de texto no estructurado en la Web:

- **Independencia del idioma:** el modelo vectorial es aplicable a toda clase de idiomas siempre que puedan extraerse palabras de los documentos (sencillo en muchos idiomas pero más complejo en otros, como el chino o el japonés, en los que no se separan las palabras).
- **Utilización únicamente de métodos estadísticos simples:** el modelo vectorial cumple totalmente este requisito al calcularse los pesos de los términos a partir de datos extraíbles directamente de los textos sin ningún tipo de conocimiento lingüístico[1].
- **Escalabilidad:** el modelo vectorial cumple este requisito parcialmente puesto que al necesitarse una fase de "indexado" en la que se generen los vectores de todos los documentos de la colección no es posible disponer de colecciones que crezcan de manera continua sino que, en ciertas ocasiones, es necesario "detener" el sistema y volver a generar toda la información del mismo (recuérdese que para calcular el peso de un término es necesario conocer en cuántos documentos se utiliza dicho término).

Por otro lado, el requisito de alta tolerancia al ruido es difícilmente alcanzable con implementaciones del modelo vectorial que utilizan palabras como términos. Puesto que palabras distintas constituyen términos distintos y, por tanto, coordenadas diferentes, los errores tipográficos, la utilización inconsistente de marcas diacríticas o simples faltas de ortografía influirán en los resultados obtenidos. Esto puede solucionarse parcialmente con algoritmos de *stemming* pero, de nuevo, se trata de conocimiento lingüístico.

3 Utilización de *n*-gramas en el modelo vectorial

En el apartado anterior se describió el modelo vectorial de manera sencilla mostrando su aplicabilidad a tareas de categorización, clasificación y recuperación de información. Se afirmó que dicho modelo, aunque aplicable a multitud de idiomas, no es excesivamente tolerante al ruido si se emplean palabras como términos. Sin embargo, el modelo no especifica qué elementos de un documento deben utilizarse como términos, es decir, no exige que se empleen palabras y admite otras posibilidades.

Una de las modificaciones más sencillas fue utilizada por Salton (1968) y consiste en utilizar no palabras sino versiones "reducidas" de las mismas tras aplicar un algoritmo de *stemming*. Esta modificación permite, en cierta medida, reducir el número de términos y "fusionar" algunos relacionados semánticamente entre sí. Sin embargo, y dejando a un lado el hecho de que hay que construir un *stemmer* para cada idioma, esta técnica no mejora el comportamiento frente al ruido (véase Fig. 22).

Una alternativa mucho más sencilla, puesto que no requiere implementar algoritmos específicos para cada idioma, y que es mucho más tolerante a errores[2] en el texto consiste en

[1] En realidad, la mayoría de implementaciones utilizan las ya mencionadas listas de "palabras vacías" que suponen el uso de conocimiento lingüístico. No es este el caso de la técnica propuesta por el autor.

[2] La utilidad de los *n*-gramas para enfrentarse a texto "ofuscado" quedó patente durante la *Confusion Track* del *TREC-5* (*Text REtrieval Conference*, Congreso sobre Recuperación de Texto). El objetivo de esa tarea era recuperar documentos para los que se disponía de versiones con ruido (5 y 20%) debido a un escaneado de baja resolución. La mayoría de participantes emplearon *n*-gramas de un modo u otro para afrontar la tarea obteniendo resultados muy satisfactorios (Kantor y Voorhees 2000).

el uso de *n*-gramas. De manera genérica, un ***n*-grama** es una secuencia de *n* elementos, palabras o caracteres, extraídos de un texto de forma no necesariamente correlativa. Sin embargo, se entiende habitualmente que un *n*-grama es una secuencia de *n* caracteres contiguos que puede contener blancos[1] y, por tanto, estar formado por segmentos de varias palabras consecutivas.

```
* busineses → busines
  businesses → busi
     busy → busi
 * desgined → desgin
  designed → design
   design → design
```

Fig. 22 Palabras inglesas procesadas con el algoritmo de *stemming* de Porter.

Se muestran en negrita aquellas palabras a las que el *stemmer* asocia la misma forma reducida, precedidas de un asterisco se presentan versiones con errores tipográficos. Obsérvese que la forma reducida asignada por el algoritmo no es la misma.

El uso de *n*-gramas en tareas de recuperación de información tiene una tradición de, al menos, 30 años. Durante este tiempo se han implementado múltiples modelos con diferencias de planteamiento muchas veces sutiles. A fin de esbozar someramente la trayectoria que se ha seguido en este campo de investigación se hará referencia a los trabajos realizados por Barton *et al.* (1974), D'Amore y Mah (1985), Cavnar (1994) y McNamee y Mayfield (2004).

```
* busineses → _bus, busi, usin, sine, ines, nese, eses, ses_
 businesses → _bus, busi, usin, sine, ines, esse, sses, ses_
                busy → _bus, busy, usy_
   * desgined → _des, desg, esgi, sgin, gine, ined, ned_
      designed → _des, desi, esig, sign, igne, ned_
         design → _des, desi, esig, sign, ign_
```

Fig. 23 4-gramas obtenidos para una serie de palabras inglesas.

Se muestran precedidas de un asterisco las versiones con errores tipográficos y en negrita aquellos 4-gramas comunes a varias de las palabras. Obsérvese cómo aquellas palabras que serían "fusionadas" por un *stemmer* comparten un gran número de *n*-gramas y que las formas incorrectas comparten varios *n*-gramas con las formas correctas. Se representa el blanco por un guión bajo.

Barton *et al.* (1974) analizaron la forma de obtener, de modo automatico, *n*-gramas de longitud variable de tal modo que su frecuencia de aparición en un ***corpus*** adecuado fuese similar: mayor o igual que un umbral fijado empíricamente (véase Fig. 24). Esto tenía como principal objetivo reducir el número de "índices" a utilizar en un diccionario de términos al tiempo que se garantizaba que dichos términos eran los más frecuentemente utilizados en los documentos[2].

La colección indexada de documentos se representaba como una matriz de bits donde cada columna representaba un documento y cada fila un término. Los bits activos indicaban la aparición del término en el correspondiente documento. Para realizar una consulta simplemente se debía obtener un vector de bits para dicha consulta y determinar qué documentos presentaban más bits en común (véase Fig. 25).

En este sentido, podría considerarse que Barton *et al.* implementaron un modelo booleano basado en *n*-gramas de longitud variable. Por otro lado, el interés fundamental de su investigación radicaba en aspectos tales como la eficiencia espacial y temporal por lo que,

[1] Algunos investigadores no sólo incluyen blancos sino cualquier tipo de símbolo de puntuación.

[2] En realidad en los títulos de los documentos, sin embargo, conceptualmente este detalle es irrelevante.

aunque constatan unos resultados satisfactorios, no hacen ninguna comparación con otros sistemas.

Corpus: Using direct access computer files of bibliographic information, an attempt is made to overcome one of the problems often associated with information retrieval, namely, the maintenance and use of large dictionaries, the greater part of which is used only infrequently.

Índices usando *n*-gramas de tamaño máximo 5 y con una frecuencia absoluta igual o superior a 3:

the	_inf	_of_	the_	tion	e_o	he_	inf	ion	n_a	of_
_a	_o	an	ar	at	d_	e_	en	er	es	f_
ic	ma	na	nf	on	re	s_	t_	te	us	_
a	b	c	d	e	f	g	h	i	l	m
n	o	p	q	r	s	t	u	v	w	y

Índices usando *n*-gramas de tamaño máximo 5 y con una frecuencia absoluta igual o superior a 4:

_of_	of_	_a	_i	_o	e_	f_	in	io	ma	n_
on	re	s_	te	th	_	a	b	c	d	e
f	g	h	i	l	m	n	o	p	q	r
s	t	u	v	w	y					

Fig. 24 Índices extraídos para un *corpus* mínimo según el método de Barton *et al.* (1974).

El objetivo básico de la técnica de Barton *et al.* (1974) era reducir el número de índices necesarios para representar documentos y consultas. Su método requiere dos parámetros: el máximo tamaño deseado para los n-gramas y la frecuencia umbral que deben superar en el *corpus* los *n*-gramas para considerarse índices. Según los propios autores este último parámetro debe determinarse de manera empírica. Nótese que el número de índices es inversamente proporcional a la frecuencia umbral seleccionada y que se incluyen entre los índices todos los caracteres individuales que aparecen en el *corpus*. En aras de la claridad se ha sustituido el espacio en blanco por el guión bajo.

Raymond D'Amore y Clinton P. Mah (1985) desarrollan un método situado a medio camino entre la propuesta de Barton *et al.* (1974) y una implementación del modelo vectorial basada en *n*-gramas de caracteres. Al igual que Barton *et al.*, su principal objetivo es reducir el número de índices para lo cual emplean *n*-gramas de caracteres aproximadamente equipobrables en un *corpus* de referencia (no necesariamente coincidente con la colección de documentos a indexar).

Consulta: An information-theoretic approach to text searching in direct access systems

Representación de la consulta usando los índices "5/3" (con información redundante):

_inf	tion	inf	ion	_a	an	ar	at	es	ic	ma
nf	on	re	s_	t_	te	-	_	a	c	d
e	f	g	h	i	m	n	o	p	r	s
t	x	y								

Representación de la consulta usando los índices "5/3" (sin información redundante):

_inf	tion	_a	an	ar	at	es	ic	ma	re	s_
t_	te	-	_	a	c	d	e	g	h	i
m	n	o	p	r	s	t	x	y		

Representación de la consulta usando los índices "5/4" (con información redundante):

_a	_i	in	io	ma	n_	on	re	s_	te	th
-	_	a	c	d	e	f	g	h	i	m
n	o	p	r	s	t	x	y			

Representación de la consulta usando los índices "5/4" (sin información redundante):

_a	_i	f_	in	io	ma	re	s_	te	th	-
_	a	c	d	e	f	g	h	i	m	n
o	p	r	s	t	x	y				

Fig. 25 Consulta representada mediante índices extraídos según el método de Barton *et al.* (1974).

Las consultas se representan empleando los índices extraídos previamente del *corpus*. La consulta puede contener información redundante (es decir, *n*-gramas que se solapan) o no. Una vez obtenida la consulta la comparación con los documentos es booleana.

D'Amore y Mah proponen utilizar conjuntos con un número fijo de índices que serían *n*-gramas de distintas longitudes aunque, fundamentalmente, emplean 2- y 3-gramas. A diferencia de Barton *et al.* no proporcionan un método automático para la obtención de tales *n*-gramas puesto que, afirman, deben determinarse experimentalmente para cada aplicación. Según estos investigadores un conjunto de índices completo (para textos en inglés) contendría aproximadamente 6.500 *n*-gramas: todos los 2-gramas alfanuméricos (36x36) junto con 200x26 3-gramas alfabéticos. Dichos 3-gramas se obtendrían "extendiendo" los 200 2-gramas alfabéticos más frecuentes en inglés y es precisamente esa selección de 3-gramas la que requiere un análisis de la colección[1]. Una vez seleccionados los índices se determina para cada *n*-grama *i* su peso w_i:

$$w_i = \frac{1}{\sqrt{p_i}} \equiv \frac{1}{\sqrt{N_i / N}} = \sqrt{\frac{N}{Ni}}$$

Donde p_i es la probabilidad de aparición del *n*-grama *i* en el *corpus* de referencia que D'Amore y Mah calculan como el cociente entre el número de documentos que contienen el *n*-grama *i*, N_i, y el número total de documentos en el *corpus*, *N*. De este modo, el peso que dichos autores asignan a cada *n*-grama del conjunto de índices es, conceptualmente, muy similar a la aplicación de *idf* (Spärck-Jones 1972) –véase ecuación en página 137.

Por su parte, cada documento es representado mediante un vector de *n*-gramas tomados del conjunto de índices a los que se asigna su frecuencia relativa de aparición en el documento. Para calcular la similitud entre dos documentos (o entre un documento y una consulta) *d* y *q* se debe calcular su producto escalar aunque introduciendo en dicho cálculo el peso fijado *a priori* para cada *n*-grama:

$$S(d,q) = \sum_x w_x \cdot f_x^d \cdot f_x^q$$

Donde w_x es el peso del *n*-grama *x* en el *corpus* de referencia y f_x^d y f_x^q son las frecuencias relativas de aparición de *x* en los documentos *d* y *q*, respectivamente. Nótese que este modo de calcular la similitud interdocumental es semejante al uso de un esquema de ponderación *tf*idf*.

Así, esta propuesta es relativamente similar a una implementación del modelo vectorial basada en *n*-gramas con las salvedades de no emplear todos los *n*-gramas posibles sino un número reducido, utilizar como medida de asociación el producto escalar en lugar de la función del coseno y aplicar un esquema de ponderación ligeramente distinto del *tf*idf* "tradicional".

Existe, no obstante, otro aspecto en que la técnica de D'Amore y Mah difiere del modelo vectorial y es el uso de un valor umbral para diferenciar las similitudes significativas de las no significativas (casuales). Para ello es necesario determinar la similitud mínima esperable en el *corpus* de referencia para tomar en consideración sólo aquellos valores de similitud que superan este valor mínimo. Dicho umbral se calcularía de este modo:

$$\sum_x w_x \cdot p_x \cdot p_x = \sum_x w_x \cdot p_x^2 = \sum_x \frac{1}{\sqrt{p_x}} \cdot p_x^2 = \sum_x p_x^{3/2}$$

[1] Por ejemplo, dado el 2-grama `th`, ¿qué 3-gramas se escogerían como índices? ¿`the`, `thy`, `ith`, …?)

Donde w_x y p_x son, respectivamente, el peso y la probabilidad del *n*-grama *x* en el *corpus* de referencia.

En resumen, D'Amore y Mah desarrollan un sistema próximo al modelo vectorial pero que emplea un conjunto fijo de 2- y 3-gramas de caracteres como términos, el producto escalar como medida de asociación, una modificación del esquema de ponderación *tf*idf* tradicional y sólo considera como significativos (no casuales) aquellos valores de similitud que superan un valor mínimo esperable. D'Amore y Mah señalan que su técnica ofrece resultados próximos a los sistemas de recuperación de información en texto completo aunque no proporcionan ninguna evaluación de dicho rendimiento.

Posteriormente, William B. Cavnar (1994) implementó un sistema fiel al modelo vectorial empleando *n*-gramas como términos en lugar de palabras o *stems*. Cavnar obtuvo resultados análogos a los de D'Amore y Mah que venían a confirmar que los *n*-gramas podían competir, en cuanto a resultados, con otros sistemas que utilizaban palabras como términos y que, a diferencia de su sistema, requerían el uso de "artefactos" como *stemming* o eliminación de palabras vacías.

Más recientemente, el sistema *HAIRCUT*[1] desarrollado en la universidad Johns Hopkins ha probado nuevamente que la utilización (sola o combinada) de técnicas "ligeras" (sin utilización de conocimientos lingüísticos previos), entre las que se incluye el uso de *n*-gramas, y métodos estadísticos pueden resultar *"al menos tan efectivos como enfoques que utilizan tratamientos dependientes del idioma y quizás más"* (McNamee y Mayfield 2004).

Así pues, puede afirmarse que la utilización de *n*-gramas para tareas de recuperación de información con independencia del modelo teórico[2] y del idioma sobre los que se apliquen proporciona resultados comparables a los obtenidos con técnicas adaptadas a cada idioma. *blindLight* enlaza con esta línea de investigación aunque, como se verá más adelante, se diferencia en algunos aspectos importantes.

Por otro lado, las aplicaciones de los *n*-gramas van más allá de la recuperación de información y, de hecho, se han aplicado a todas las tareas mencionadas al comienzo del capítulo, específicamente a la categorización y clasificación de documentos, así como a la extracción automática de términos clave.

En este sentido requieren mención especial los trabajos realizados por Marc Damashek (1995) y Jonathan D. Cohen (1995) el primero con la técnica *Acquaintance* y el segundo con *Highlights*. Puesto que la nueva técnica propuesta por el autor de este trabajo establece, en cierta medida, un puente entre las desarrolladas por Cohen y Damashek es necesario describir éstas brevemente antes de poder presentar los fundamentos de *blindLight* y señalar las diferencias entre la propuesta del autor y el resto de métodos desarrollados hasta la fecha.

3.1 Estimación de la similitud interdocumental utilizando *n*-gramas (*Acquaintance*)

La propuesta de Damashek (1995) presenta semejanzas con la de D'Amore y Mah (1985). Al igual que ellos, *Acquaintance* representa los documentos como vectores de pesos de *n*-gramas donde cada peso es la frecuencia relativa de aparición del correspondiente

[1] `http://haircut.jhuapl.edu/index.html`

[2] Es posible desarrollar sistemas de recuperación de información que usen *n*-gramas como términos y que implementen el modelo vectorial, el probabilístico o cualquier otro tipo de modelo *ad hoc*. En este sentido la aplicación actual de *blindLight* a esta tarea, aunque similar en ciertos aspectos al modelo vectorial, debería considerarse como un modelo diferente.

n-grama en el documento y también utiliza el producto escalar como métrica para comparar los vectores.

No obstante, Damashek no utiliza *n*-gramas de tamaño variable ni tampoco establece un conjunto de *n*-gramas de tamaño fijo para realizar el indexado. Además, Damashek no sólo emplea la técnica *Acquaintance* para llevar a cabo recuperación de información sino que la extiende para realizar categorización y clasificación de documentos. Este paso resulta natural puesto que, como se mostró en el apartado dedicado al modelo vectorial, estas tareas pueden llevarse a cabo fácilmente si se dispone de una métrica de la similitud entre documentos y de un modo de obtener centroides para conjuntos de documentos, ambas cosas posibles con *Acquaintance*.

Hay que señalar, sin embargo, dos debilidades en esta técnica. En primer lugar, el rendimiento obtenido al realizar recuperación de información empleando consultas cortas (las más habituales en la Web) es, en palabras del propio Damashek, pobre. *Acquaintance* ofrece, en cambio, mejores resultados cuando se recupera información partiendo de un documento de ejemplo. En segundo lugar, Damashek (1995) afirma también lo siguiente:

> *La métrica [de similitud entre documentos] falla en tareas más sutiles como la discriminación basándose en el asunto [del documento] puesto que los vectores obtenidos a partir de texto plano están habitualmente dominados por componentes no-informativas (por ejemplo, en inglés los n-gramas derivados de "is the", "and the", "with the", …) y son estos componentes de mayor peso los que más influyen en el producto escalar.*

Para enfrentarse a este problema Damashek sugiere realizar una traslación del conjunto de documentos sobre los que van a realizarse las medidas de similitud a fin de minimizar la influencia de estos componentes "no-informativos". Para ello, propone sustraer, componente a componente, el centroide de cada elemento del conjunto. Esto puede realizarse para toda la colección de documentos (lo que equivaldría en cierta medida a eliminar palabras vacías) o para subconjuntos de la misma obtenidos tras una clasificación.

En resumen, *Acquaintance* es una implementación del modelo vectorial en la que se obtiene un vector de pesos de *n*-gramas para cada documento de la colección. Dichos pesos no son más que la frecuencia relativa de aparición de cada *n*-grama en el documento en cuestión. Para evaluar la similitud entre dos vectores se calcula el producto escalar de los mismos y a fin de evitar la influencia de componentes de peso elevado pero poca o nula capacidad discriminativa se traslada el conjunto de documentos sustrayendo de cada vector el centroide de la colección. De este modo, es posible llevar a cabo recuperación de información empleando documentos de ejemplo así como clasificación y categorización.

3.2 Extracción automática de términos clave utilizando *n*-gramas (*Highlights*)

El objetivo fundamental de *Highlights* (Cohen 1995) es extraer de manera automática unos pocos términos clave[1] que permitan a un usuario determinar con rapidez el asunto tratado en un documento. Esta técnica permitiría, por ejemplo, decidir qué documentos obtenidos como respuesta a una consulta son verdaderamente relevantes y cuáles no. Cohen, al igual que el autor de este trabajo, estableció una serie de requisitos para su técnica: debía ser independiente del lenguaje, del dominio de conocimiento y estar basada tan sólo en métodos estadísticos. Para ello, utilizó *n*-gramas de caracteres y vectores de *n*-gramas para representar tanto los documentos como la colección.

[1] Es necesario notar que los términos clave podían ser no sólo palabras sino frases (por ejemplo, `recuperación de información`, `interfaz de usuario`, `lenguajes de programación`, etc.)

Highlights utiliza vectores de pesos de *n*-gramas de manera similar a como se utilizaban en *Acquaintance*. No obstante, como fase previa al cálculo de los vectores de los distintos documentos era necesario construir un "contexto" dentro del cual se realizaría el posterior procesamiento. Dicho contexto no era más que un vector obtenido al tratar toda la colección como un único texto de gran longitud[1]. Los vectores de los distintos documentos eran comparados entonces con este contexto para determinar qué *n*-gramas de cada documento eran menos "probables" con relación a la colección o lo que es lo mismo, menos comunes y, por tanto, más interesantes para describir el documento en cuestión.

Finalmente, una vez establecidos los pesos para cada *n*-grama de un documento *Highlights* procedía a "puntuar" los caracteres del texto. Para ello cada vez que un *n*-grama aparecía en el documento se asignaba su puntuación al carácter central del mismo (véase Fig. 26). Al finalizar esta fase se establecía un umbral que separaba los caracteres "interesantes" de los "no interesantes". Mediante este umbral se extraían aquellas palabras o frases que incluyesen algún carácter relevante. Estos términos eran incluidos en la lista final de resultados que sólo necesitaba ser ordenada antes de ser ofrecida al usuario.

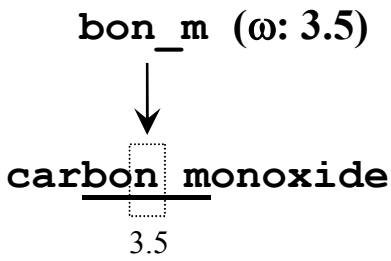


Fig. 26 El peso de un *n*-grama es asignado al carácter central del mismo.

Highlights "puntua" los caracteres que aparecen en un documento a fin de localizar los términos clave. Para ello se localizan las apariciones de cada *n*-grama en el texto y se asigna su peso al carácter central del *n*-grama.

4 Obtención de resúmenes automáticos

Al comienzo del capítulo se señaló que el problema de procesar texto en la Web podía dividirse en un conjunto de tareas tales como: categorización, clasificación, recuperación de información y extracción de información. En los apartados anteriores se han analizado las características fundamentales de las tres primeras tareas y se ha visto que, debido a su estrecha relación, todas pueden ser resueltas empleando prácticamente las mismas técnicas.

Por otro lado, teniendo en cuenta las características del texto disponible libremente en la Web se puso énfasis en aquellas técnicas que mostrasen una alta tolerancia al ruido y empleasen métodos estadísticos simples. Así, se vio cómo la utilización de *n*-gramas[2] resulta particularmente interesante.

Se mostraron seguidamente dos técnicas especialmente relevantes basadas en el uso de tales *n*-gramas. La primera, *Acquaintance* (Damashek 1995), facilitaba la comparación de documentos permitiendo la categorización, clasificación y recuperación de información. La

[1] La idea guarda similitudes con el uso de centroides en *Acquaintance* pero su construcción es totalmente diferente.

[2] Entendidos éstos como secuencias de caracteres, espacios en blanco incluidos, extraídos del texto de manera correlativa.

segunda, *Highlights* (Cohen 1995), tenía como objetivo la extracción de una serie de términos clave (palabras e incluso frases) que permitiesen que un usuario determinase con facilidad el asunto tratado en un documento diferenciándolo del resto de documentos de su entorno.

Esta última es la más próxima a la cuarta tarea de interés para aliviar la sobrecarga de información en la Web: el resumen automático. Sin embargo, no puede considerarse en modo alguno que la resuelva ya que *Highlights* se limita a proporcionar una lista ordenada de términos relevantes para un documento y nunca una versión resumida del mismo.

Puesto que uno de los objetivos de la nueva técnica presentada en este trabajo es la obtención automática de un resumen a partir de un único documento[1], en este apartado se describirán a grandes rasgos los aspectos más relevantes del campo. Se ofrecerán más detalles en el capítulo 7.

Luhn, al que ya se citó como uno de los pioneros en el campo del tratamiento automático de textos, fue el primero en proponer un sistema para obtener un resumen de un documento empleando medios mecánicos (Luhn 1958). Los resúmenes construidos por su sistema eran extractivos, genéricos e informativos[2]. Extractivos puesto que el resumen consistía en una selección del texto original. Genéricos al construírse siempre del mismo modo, reflejando el punto de vista del autor del documento sin permitir que el usuario los orientase para satisfacer posibles consultas. E informativos puesto que incluían los contenidos más relevantes del original en lugar de describir la naturaleza del texto.

No obstante, es posible obtener resúmenes por abstracción, esto es, el sistema no extrae sentencias del texto original sino que "redacta" un documento completamente nuevo. Asimismo, los resúmenes pueden adaptarse a los requisitos que el usuario especifique en forma de consulta. O pueden ser descriptivos, es decir, sin reflejar los contenidos del documento original pueden indicar la naturaleza del mismo.

Es necesario decir que el resumen automático por abstracción así como la construcción de resúmenes descriptivos requieren de técnicas mucho más sofisticadas que para la "simple" extracción de información relevante; de hecho *"no es probable que se construyan sistemas prácticos de resumen por abstracción en el futuro cercano"* (Hovy 1999, p. 7). Por otro lado, generar resúmenes adaptados al usuario siempre puede afrontarse como una tarea de recuperación de información en la que cada sentencia del documento es tratada como un documento individual.

A la hora de afrontar la tarea de obtener resúmenes automáticos el autor de este trabajo se ha centrado únicamente en la construcción de resúmenes extractivos, genéricos e informativos. Por otro lado, como ya se dijo con anterioridad y se verá más adelante, la nueva técnica propuesta, *blindLight*, utiliza *n*-gramas de caracteres no sólo para las tareas de categorización, clasificación y recuperación de información sino también para la construcción de los resúmenes. Es por ello que se afirmaba que establecía un vínculo entre *Acquaintance* (Damashek 1995) y *Highlights* (Cohen 1995).

Hasta donde sabe el autor, tan sólo ha habido hasta la fecha otro intento de obtener resúmenes extractivos empleando *n*-gramas de caracteres. Joel Larocca Neto *et al.* (2000) construyeron un sistema para generar resúmenes extractivos adaptando la técnica de ponderación *tf*idf* a colecciones de sentencias en lugar de colecciones de documentos. Así,

---

[1] Es decir, el único *"corpus"* utilizado para la obtención de información estadística acerca del lenguaje utilizado es el propio documento a resumir.

[2] Las definiciones para resúmenes por extracción/abstracción, informativo/descriptivo y genérico/adaptado al usuario son las recogidas por Eduard Hovy (1999).

proponían una nueva medida, *tf*isf*, para ponderar los términos de un texto donde *isf* es el número de sentencias que incluyen un término dado. El resumen se construye con aquellas sentencias con un valor *tf*isf* más elevado. Los *n*-gramas de caracteres son uno de los dos tipos de términos que pueden utilizarse en su sistema, siendo el segundo palabras individuales una vez eliminadas aquellas vacías de contenido y aplicado un *stemmer*.

Así pues, la novedad de la propuesta de Neto *et al.* es relativa puesto que la utilización de *n*-gramas como términos de modelos vectoriales es bien conocida y ya se ha descrito en apartados anteriores. Por otro lado, cuando se describa a lo largo del resto de la disertación la técnica *blindLight* se podrá comprobar que la semejanza entre ambas propuestas se limita a la utilización de *n*-gramas de caracteres puesto que la forma en que se calculan los pesos de dichos *n*-gramas, se utilizan para representar los documentos, se calculan las similitudes, así como la manera en que se construyen los resúmenes son totalmente diferentes.

CAPÍTULO 3

DESCRIPCIÓN DE LA TÉCNICA *BLINDLIGHT*

En el capítulo anterior se analizaron algunas de las técnicas aplicables a tareas de PLN como clasificación, categorización, recuperación de información o extracción de resúmenes poniendo énfasis en los métodos estadísticos y en aquellos aspectos "transversales"a las distintas técnicas (p.ej. la representación vectorial de los documentos o el concepto de similitud interdocumental). En este capítulo se describirá blindLight, la técnica propuesta como prueba empírica de la tesis del autor. Se trata de un método de PLN de inspiración biológica, sencillo, robusto, y aplicable a múltiples idiomas. La idea subyacente es simple: obtener para cada documento un "genoma" único e independiente de cualquier colección que incluya a dicho documento. Este genoma estará constituido por n-gramas, secuencias de unos pocos caracteres extraídos del texto y ponderados estadísticamente para indicar su distinto grado de significatividad. Como se verá, blindLight está relacionada con el modelo vectorial y, al igual que éste, puede ser aplicada a la clasificación, categorización y recuperación de documentos aunque presenta importantes diferencias respecto al mismo. Por otro lado, la idea de un "genoma" documental es más que una metáfora puesto que dicho genoma puede ser "activado" resumiendo el documento del cual fue "extraído". A lo largo de este capítulo se describirá la técnica blindLight comparándola con otras técnicas aplicables a tareas similares y se apoyará el argumento del autor acerca de la capacidad de esta técnica para representar la semántica subyacente a los textos con independencia del idioma en que estén escritos.

1 *blindLight*, una técnica bio-inspirada

La aplicación de técnicas bioinformáticas al tratamiento del lenguaje natural como propuso el autor para la Web Cooperativa no es una idea nueva. La alineación de múltiples secuencias de texto se ha aplicado, por ejemplo, a la generación de texto (Barzilay y Lee 2002), al aprendizaje automático de técnicas de paráfrasis (Barzilay y Lee 2003) o a la inducción de gramáticas (Kruijff 2002).

Tampoco es nueva la idea de "extraer" algún tipo de "ADN" a partir de texto libre. Por ejemplo, la mayor parte de las tecnologías pendientes de patente y presuntamente

desarrolladas por la empresa *Meaningful Machines*[1] emplean de un modo u otro la idea de "ADN del lenguaje":

> *existe un número finito de ideas discretas […] que Fluent Machines llama el 'ADN' del significado y que son universales y expresables en cualquier idioma.* (Abir *et al.* 2002, p. 216)

Resulta claro, en especial tras estudiar las solicitudes de patente (Abir 2003a, 2003b, 2003c y 2004), que su sistema se basa en el uso de *n*-gramas de caracteres y textos paralelos para establecer la asociación entre *n*-gramas de distintos idiomas:

> *el sistema […] construye bases de datos multilingües que contienen n-gramas de ADN de diversas longitudes […] y conecta traducciones de n-gramas en el lenguaje objetivo produciendo texto traducido.* (Abir *et al.* 2002, p. 217)

Puede parecer entonces que *blindLight* no es una técnica excesivamente novedosa al pretender utilizar técnicas bioinformáticas (básicamente alineación de secuencias) para comparar cadenas de pseudo-ADN constituidas, a su vez, por *n*-gramas de caracteres. No obstante, como se irá mostrando a lo largo de este trabajo, *blindLight* supone la aportación de una serie de ideas nuevas y originales:

1. Para construir el "genoma" de los documentos se emplea una medida estadística de la **"significatividad"** de los distintos *n*-gramas que va un paso más allá de la clásica frecuencia de aparición.
2. El "genoma" de un documento puede "actuar" sobre el texto del documento, a modo de "ARN transferente", transformándolo en resúmenes y frases clave.
3. No se emplean técnicas de alineación de secuencias para las comparaciones de "genomas" sino que éstos pueden combinarse constituyendo "híbridos" que serán comparados con los originales a fin de determinar la similitud entre los mismos.

Una versión preliminar de estas ideas, en particular de las dos primeras, se puede encontrar en (Gayo Avello, Álvarez Gutiérrez y Gayo Avello 2004a y 2004b). En el siguiente apartado se describirá con detalle cómo se construyen y comparan tales "genomas documentales" y en posteriores capítulos el modo en que dichos "genomas" transforman el texto plano original generando resúmenes automáticos. Baste para concluir la definición de "ADN de un documento" tal y como se entiende en *blindLight*:

> *El ADN de un documento es un conjunto de genes donde cada gen está formado por un n-grama de caracteres y su correspondiente significatividad dentro del documento de origen.*

2 Fundamentos teóricos de *blindLight*

En este apartado se describirá con detalle la nueva técnica que satisface la afirmación con que el autor comenzaba su tesis:

> *Se puede obtener para los distintos n-gramas, g_i, de un texto escrito en cualquier idioma una medida de su significatividad, s_i, distinta de la frecuencia relativa de aparición de los mismos en el texto, f_i, pero calculable a partir de la misma. Esta métrica de la significatividad intradocumental de los n-gramas permite asociar a cada documento, d_i, un único vector, v_i, susceptible de comparación con cualquier otro vector obtenido del mismo modo aun cuando sus respectivas longitudes puedan diferir.*

[1] `http://www.meaningfulmachines.com/`

blindLight, al igual que otras técnicas basadas en vectores de *n*-gramas, representa cada documento como un vector de pesos. Sin embargo, estos vectores difieren en varios aspectos de los utilizados en modelos vectoriales "clásicos". En primer lugar, no se considera a los documentos vectores en un espacio *T*-dimensional sino que dos vectores cualesquiera tendrán, muy probablemente, distintas dimensiones; o lo que es lo mismo, en esta propuesta no existe un auténtico "espacio vectorial" y no se recurre a medidas de asociación análogas a operaciones vectoriales. Por otro lado, los pesos empleados en cada vector no son las frecuencias relativas de aparición[1] de los *n*-gramas sino la "significatividad" de cada *n*-grama dentro del documento. Dicha significatividad, como se verá más adelante, se obtiene a partir de las frecuencias relativas.

Calcular una medida de la relación entre los elementos de un *n*-grama y, así, la relevancia del *n*-grama en su conjunto, su significatividad[2], no es un problema reciente. No obstante, tan sólo se citarán aquí dos trabajos representativos, el hecho por Dunning y el llevado a cabo por Ferreira da Silva y Pereira Lopes. Ted Dunning (1993) describió un método basado en el test de razón de verosimilitud (*likelihood ratio test*) para detectar palabras clave y terminología. No obstante, empleando su técnica sólo se podían detectar bigramas de palabras (por ejemplo, `likelihood ratio` o `ratio test`, nunca `likelihood ratio test`). Fueron Joaquim Ferreira da Silva y Gabriel Pereira Lopes (1999) quienes presentaron un método para generalizar una serie de estadísticos[3] a *n*-gramas de palabras de longitud arbitraria a fin de extraer frases clave. Además de esto, introdujeron una nueva medida (véase la ecuación 2), la Probabilidad Condicional Simétrica (*Symmetrical Conditional Probability*), que según sus autores supera a las anteriores, incluyendo los resultados alcanzados por Dunning.

blindLight aplica estos estadísticos no a *n*-gramas de palabras sino de caracteres, midiendo de este modo la relación entre los caracteres constituyentes de cada *n*-grama y, por tanto, la significatividad de éste dentro de un único documento. Así, para cada *n*-grama se calcula su significatividad constituyendo cada par `(n-grama, significatividad)` un componente[4] del vector correspondiente a un documento dado. Por tanto, los vectores de los documentos no se construyen en relación a un *corpus* ponderando los términos en función de la frecuencia de aparición en la colección de documentos.

La técnica no obliga a emplear un estadístico en particular para la ponderación de los *n*-gramas y debería estudiarse para cada aplicación cuál resulta el más adecuado. No obstante, es preciso señalar que, mientras no se indique lo contrario, los resultados descritos en este trabajo se han obtenido empleando la información mutua (véase la ecuación 3) y en

---

[1] Normalizadas sobre la base de un *corpus* y/o la longitud del documento.

[2] A lo largo de este trabajo se utilizará el término "significatividad" para referirse al valor real asignado a cada *n*-grama de caracteres dentro de un documento. Se usa éste término en lugar del habitual "peso" para distinguir el modo en que se obtienen ambos valores en *blindLight* y en el modelo vectorial, respectivamente. En este último el cálculo de los pesos involucra no sólo la frecuencia de aparición de los términos en el propio documento (*tf*) sino también su distribución en los distintos documentos de la colección (*idf*) siendo de hecho más importante este último valor. Por el contrario, en *blindLight* el valor asignado a cada *n*-grama se deriva únicamente a partir del propio documento siendo innecesario recurrir a la colección de documentos.

[3] Información mutua (véase ecuación 3), ϕ^2, *log likelihood* (Dunning 1993) y Dice. Las ecuaciones para la generalización a *n*-gramas del resto de estadísticos se encuentran en la página 148.

[4] Un "gen" del "ADN documental".

algunos casos, como en el sistema de recuperación de información participante en *CLEF*[1] *2004* (Peters *et al.* 2005), la probabilidad condicional simétrica (véase la ecuación 2).

$$Avp = \frac{1}{n-1}\sum_{i=1}^{i=n-1} p(w_1...w_i)\cdot p(w_{i+1}...w_n) \quad (1)$$

$$SCP\_f((w_1...w_n)) = \frac{p(w_1...w_n)^2}{Avp} \quad (2)$$

$$SI\_f((w_1...w_n)) = \log(\frac{p(w_1...w_n)}{Avp}) \quad (3)$$

Fig. 27 Cálculo de la probabilidad condicional simétrica (*SCP_f*) y la información mutua (*SI_f*) para *n*-gramas de caracteres.

$(w_1...w_n)$ es un *n*-grama, *$(w_1...w_i)$* y *$(w_{i+1}...w_n)$* son fragmentos consecutivos del mismo (p.ej. para el *n*-grama *'info'* se tendría <*'i'*, *'nfo'*>, <*'in'*, *'fo'*> y <*'inf'*, *'o'*>). *$p((w_1...w_n))$* es la probabilidad del *n*-grama *$(w_1...w_n)$* en el texto, *$p((w_1...w_i))$* es la probabilidad de que un *n*-grama comience con los caracteres *$(w_1...w_i)$* y *$p(w_{i+1}...w_n)$* de que termine en *$(w_{i+1}...w_n)$*.

A continuación se muestra el modo en que se calcula la significatividad de 4-gramas de caracteres utilizando una de las historias más cortas jamás escritas: *"El Dinosaurio"* de Augusto Monterroso.

```
Cuando despertó, el dinosaurio todavía estaba allí.
```

Fig. 28 *"El Dinosaurio"* de Augusto Monterroso.

Cuan	uand	ando	ndo_	do_d	o_de	**_des**	desp
espe	sper	pert	ertó	rtó_	tó_e	ó_el	**_el_**
el_d	l_di	**_din**	dino	inos	**nosa**	osau	saur
auri	urio	rio_	io_t	o_to	**_tod**	**toda**	odav
daví	**avía**	vía_	ía_e	a_es	**_est**	**esta**	stab
taba	aba_	**ba_a**	a_al	**_all**	allí		

Fig. 29 4-gramas del texto anterior (se han sustituido los espacios en blanco por guiones bajos)**.**

_→	**6**	_a→	1	_al→	1	_d→	2	_de→	1	_di→	1	_e→	2	_el→	1
_es→	1	_t→	1	_to→	1	a→	7	a_→	2	a_a→	1	a_e→	1	ab→	1
aba→	1	al→	1	all→	1	an→	1	and→	1	au→	1	aur→	1	av→	1
aví→	1	b→	1	ba→	1	ba_→	1	C→	1	Cu→	1	Cua→	1	d→	4
da→	1	dav→	1	de→	1	des→	1	di→	1	din→	1	do→	1	do_→	1
e→	4	el→	1	el_→	1	er→	1	ert→	1	es→	2	esp→	1	est→	1
i→	2	í→	1	ía→	1	ía_→	1	in→	1	ino→	1	io→	1	io_→	1
l→	1	l_→	1	l_d→	1	lí→	0	ll→	0	lli→	0	n→	2	nd→	1
ndo→	1	no→	1	nos→	1	o→	4	ó→	1	o_→	2	ó_→	1	o_d→	1
ó_e→	1	o_t→	1	od→	1	oda→	1	os→	1	osa→	1	p→	1	pe→	1
per→	1	r→	2	ri→	1	rio→	1	rt→	1	rtó→	1	s→	3	sa→	1
sau→	1	sp→	1	spe→	1	st→	1	sta→	1	t→	3	ta→	1	tab→	1
to→	1	tó→	1	tó_→	1	tod→	1	u→	2	ua→	1	uan→	1	ur→	1
uri→	1	v→	1	ví→	1	vía→	1								

Fig. 30 Fragmentos iniciales de los 4-gramas anteriores junto con sus frecuencias absolutas.

[1] *Cross Language Evaluation Forum* es un foro internacional que tiene como objetivos el desarrollo de una infraestructura para la evaluación de sistemas de recuperación de información que operen sobre idiomas europeos así como la creación de conjuntos de prueba reutilizables <http://www.clef-campaign.org>.

→_	6	→_a	1	→_al	1	→_d	2	→_de	1	→_di	1	→_e	2	→_el	1
→_es	1	→_t	1	→_to	1	**→a**	**6**	→a_	2	→a_a	1	→a_e	1	→ab	1
→aba	1	→al	1	→all	1	→an	1	→and	1	→au	1	→aur	1	→av	1
→aví	1	→b	1	→ba	1	→ba_	1	→C	0	→Cu	0	→Cua	0	→d	4
→da	1	→dav	1	→de	1	→des	1	→di	1	→din	1	→do	1	→do_	1
→e	4	→el	1	→el_	1	→er	1	→ert	1	→es	2	→esp	1	→est	1
→i	2	→í	2	→ía	1	→ía_	1	→in	1	→ino	1	→io	1	→io_	1
→l	3	→l_	1	→l_d	1	→lí	1	→ll	1	→llí	1	→n	2	→nd	1
→ndo	1	→no	1	→nos	1	→o	4	→ó	1	→o_	2	→ó_	1	→o_d	1
→ó_e	1	→o_t	1	→od	1	→oda	1	→os	1	→osa	1	→p	1	→pe	1
→per	1	→r	2	→ri	1	→rio	1	→rt	1	→rtó	1	→s	3	→sa	1
→sau	1	→sp	1	→spe	1	→st	1	→sta	1	→t	3	→ta	1	→tab	1
→to	1	→tó	1	→tó_	1	→tod	1	→u	1	→ua	0	→uan	1	→ur	1
→uri	1	→v	1	→ví	1	→vía	1								

Fig. 31 Fragmentos finales de los 4-gramas anteriores junto con sus frecuencias absolutas.

$$p(\text{Cuan}) = \frac{1}{46} \left| \begin{array}{lll} p(\text{C}\rightarrow) = \frac{1}{138} & p(\text{Cu}\rightarrow) = \frac{1}{138} & p(\text{Cua}\rightarrow) = \frac{1}{138} \\ \times & \times & \times \\ p(\rightarrow\text{uan}) = \frac{1}{138} & p(\rightarrow\text{an}) = \frac{1}{138} & p(\rightarrow\text{n}) = \frac{2}{138} \end{array} \right.$$

$$\frac{1}{138^2} + \frac{1}{138^2} + \frac{2}{138^2} = \frac{4}{138^2}$$

$$Avp = \frac{1}{3} \cdot \frac{4}{138^2}$$

$$SI\_f(\text{Cuan}) = \log \frac{\frac{1}{46}}{\frac{1}{3} \cdot \frac{4}{138^2}} = 2{,}492$$

Fig. 32 Cálculo de la significatividad del 4-grama `Cuan` empleando información mútua (*SI_f*).

$$p(\text{ando}) = \frac{1}{46} \left| \begin{array}{lll} p(\text{a}\rightarrow) = \frac{7}{138} & p(\text{an}\rightarrow) = \frac{1}{138} & p(\text{and}\rightarrow) = \frac{1}{138} \\ \times & \times & \times \\ p(\rightarrow\text{ndo}) = \frac{1}{138} & p(\rightarrow\text{do}) = \frac{1}{138} & p(\rightarrow\text{o}) = \frac{4}{138} \end{array} \right.$$

$$\frac{7}{138^2} + \frac{1}{138^2} + \frac{4}{138^2} = \frac{12}{138^2}$$

$$Avp = \frac{1}{3} \cdot \frac{12}{138^2}$$

$$SI\_f(\text{ando}) = \log \frac{\frac{1}{46}}{\frac{1}{3} \cdot \frac{12}{138^2}} = 2{,}015$$

Fig. 33 Cálculo de la significatividad del 4-grama `ando` empleando información mútua (*SI_f*).

Por otro lado, *blindLight* no utiliza operaciones vectoriales para comparar vectores de documentos. En cambio, para determinar la similitud entre dos vectores dados se obtiene un nuevo vector mediante la intersección de los dos anteriores y se compara la significatividad total del vector resultante con la de los vectores originales.

Esta operación es, en cierto modo, similar al coeficiente de solapamiento propuesto por Salton (1968, citado por Jardine y van Rijsbergen 1971) con dos salvedades. En primer lugar, no se interpreta la intersección entre vectores con pesos reales como un producto escalar sino como la suma de los pesos de un vector que tiene como componentes los *n*-gramas presentes en ambos documentos y como pesos los mínimos que aparecen en cada

vector original. En segundo lugar, *blindLight* no utiliza un único coeficiente sino dos al comparar el vector intersección resultante con cada vector documento de manera independiente.

El motivo de todo esto se explicará más adelante ya que antes de continuar es preferible expresar lo anterior en forma de ecuaciones. Sean Q y T dos vectores *blindLight* de dimensiones m y n:

$$Q = \{(k_{1Q}, w_{1Q}) \quad (k_{2Q}, w_{2Q}) \quad \ldots \quad (k_{mQ}, w_{mQ})\} \tag{4}$$

$$T = \{(k_{1T}, w_{1T}) \quad (k_{2T}, w_{2T}) \quad \ldots \quad (k_{nT}, w_{nT})\} \tag{5}$$

k_{ij} es el n-grama i-ésimo en el documento j y w_{ij} es la significatividad de dicho n-grama calculada aplicando cualquiera de los estadísticos generalizados a n-gramas (Ferreira da Silva y Pereira Lopes 1999).

Es posible definir entonces la significatividad total para los vectores Q y T, S_Q y S_T respectivamente, como:

$$S_Q = \sum_{i=1}^{m} w_{iQ} \tag{6}$$

$$S_T = \sum_{i=1}^{n} w_{iT} \tag{7}$$

El operador de intersección antes mencionado, Ω, se define de la forma siguiente:

$$Q\Omega T = \left\{ (k_x, w_x) \middle/ \begin{array}{l} (k_x = k_{iQ} = k_{jT}) \wedge (w_x = \min(w_{iQ}, w_{jT})), \\ (k_{iQ}, w_{iQ}) \in Q, 0 \le i < m, \\ (k_{jT}, w_{jT}) \in T, 0 \le j < n \end{array} \right\} \tag{8}$$

De manera similar a como muestran las ecuaciones 6 y 7 se puede calcular la significatividad total para el vector resultante de la intersección de los vectores a comparar:

$$S_{Q\Omega T} = \sum w_{iQ\Omega T} \tag{9}$$

Finalmente, se definen dos medidas de asociación, una para comparar Q con T, Π (Pi mayúscula), y otra para comparar T con Q, P (Ro mayúscula):

$$\Pi = S_{Q\Omega T} / S_Q \tag{10}$$

$$\mathrm{P} = S_{Q\Omega T} / S_T \tag{11}$$

Si se supone que T es un documento y Q una consulta pueden asimilarse las medidas Π y P con la precisión y la **exhaustividad** (*recall*), respectivamente. Recordemos que la precisión es la proporción de documentos retornados por un sistema que son realmente relevantes mientras que la exhaustividad es la proporción de documentos relevantes en la colección que aparecen entre los resultados. En este sentido, Π revelaría en

qué medida la consulta queda "satisfecha" por la intersección entre ésta y el documento resultante mientras que P señalaría lo propio entre la intersección y el documento.

De este modo, un valor de Π igual a la unidad implicaría que la necesidad de información formulada en la consulta queda totalmente satisfecha con el documento mientras que un valor de P unitario indicaría que la consulta define por completo al documento. Naturalmente, puesto que la significatividad total de consulta y documento depende en gran medida del número de *n*-gramas de cada uno y, en consecuencia, de la longitud del texto, los valores de Π y P difícilmente tomarán valores próximos a la unidad de manera simultánea[1] por lo que será necesario combinar ambas medidas de modo que sea posible obtener un único valor que indique la similitud entre documentos de distinto tamaño (para más detalles sobre este tipo de medidas véase la página 141). A continuación se presenta un ejemplo ilustrativo de estos conceptos empleando de nuevo el texto de Monterroso junto con su traducción al portugués.

Cuando despertó, el dinosaurio todavía estaba allí. (*Q*)
Quando acordou, o dinossauro ainda estava lá. (*T*)

Fig. 34 *"El Dinosaurio"* de Augusto Monterroso, original en español y traducción al portugués.

Vector Q (45 elementos)		**Vector T (39 elementos)**		**QΩT (10 elementos)**	
Cuan	2,492	va_l	2,545	saur	2,244
l_di	2,392	rdou	2,323	inos	2,177
stab	2,392	stav	2,323	uand	2,119
…		…		_est	2,091
saur	2,313	saur	2,244	dino	2,022
desp	2,313	noss	2,177	_din	2,022
…		…		esta	2,012
ndo_	2,137	a_lá	2,022	ndo_	1,981
nosa	2,137	o_ac	2,022	a_es	1,943
…		…		ando	1,876
ando	2,015	auro	1,908		
avía	1,945	ando	1,876		
_all	1,915	do_a	1,767	**Π: 0,209 P: 0,253**	

Fig. 35 Vectores *blindLight* para los documentos mostrados en Fig. 34.

Los vectores se han calculado siguiendo el proceso descrito desde Fig. 29 a Fig. 33, además, se han truncado para mostrar 10 elementos (a excepción del vector intersección que aparece completo). Los espacios en blanco han sido reemplazados por guiones bajos.

Es preciso señalar que los vectores *blindLight* no están normalizados, esto es, su módulo no es unitario (véase Fig. 35) sino que los valores de significatividad obtenidos en cada documento son utilizados directamente como pesos de los *n*-gramas. La razón es sencilla, puesto que el operador intersección antes definido produce un nuevo vector que tiene como pesos los mínimos de los vectores intersecados, la normalización de los mismos tendría como consecuencia que el documento más largo de los dos, aquel, por tanto, con más *n*-gramas diferentes y, en consecuencia, menores pesos sería siempre el más influyente en la comparación lo cual es contrario al objetivo de la normalización, esto es, garantizar que el tamaño de los documentos no es un factor determinante en las comparaciones.

[1] Salvo que se comparen documentos de tamaño similar como en los siguientes ejemplos.

Por otro lado, si todos los vectores fuesen unitarios no tendría sentido alguno hablar de dos medidas de comparación asimétricas que, en opinión del autor, dotan de gran flexibilidad a la nueva técnica ya que combinando linealmente Π y P es posible construir nuevas medidas de asociación. En el capítulo dedicado a recuperación de información se tratará más este asunto, baste mostrar aquí una de las más simples[1], la denominada *PiRo*:

$$\frac{\Pi + P}{2} \qquad (\mathbf{12})$$

Esta medida de asociación es, por supuesto, simétrica[2] y proporciona valores entre 0 y 1 (parecido nulo y documentos idénticos, respectivamente). Pruebas experimentales manifiestan una correlación elevada entre esta medida de asociación y la función del coseno; podría ser interesante verificar si es o no monótona con respecto a esa y otras funciones de asociación pero, no siendo central para este trabajo, aún no se ha llevado a cabo dicho estudio. Por otro lado, y adelantándose al capítulo dedicado a trabajo futuro, la utilización de dos medidas asimétricas como base para construir medidas de asociación ofrece la interesante posibilidad de utilizar programación genética para obtener nuevas medidas de un modo similar al seguido por Fan, Gordon y Pathak (2004a y 2004b).

Además, existe una serie de trabajos relativos a nuevas métricas de la similitud entre *ítems* de información que resultan muy interesantes de cara a una futura integración con la técnica del autor:

- Bennett *et al.* (1998) describen cómo se puede elaborar una medida universal de la distancia "cognitiva" entre dos objetos cualesquiera (representables mediante cadenas de bits) basándose en la complejidad de Kolmogorov (función *K*). Puesto que dicha función *K* no es computable Chen, Kwong y Li (1999) la aproximan a partir de los resultados obtenidos con un algoritmo de compresión[3] para cadenas de ADN sugiriendo la posibilidad de emplear esta técnica para la comparación de genomas sin necesidad de recurrir a su alineación. Posteriormente, Li *et al.* (2004) continuan los trabajos anteriores demostrando su aplicación a la clasificación de organismos biológicos y lenguajes naturales. Varré, Delahaye y Rivals (1999) llevaron a cabo un trabajo muy similar[4].
- Rubner, Tomasi y Guibas (2000) describen la *Earth Mover's Distance* (*EMD*) basada en el coste mínimo necesario para transformar una distribución de valores en otra. Algunas de las ventajas de la *EMD* son su capacidad para trabajar sobre representaciones de longitud variable y para soportar coincidencias parciales. Por el momento, esta distancia se ha utilizado con imágenes, no con texto.
- Muthukrishnan y Sahinalp (2000) estudian el problema de los "vecinos más próximos a una secuencia" (*sequence nearest neighbors*), es decir, la forma de encontrar en una colección *D* de secuencias aquella cuya distancia de edición a otra secuencia

[1] Como se describirá en el capítulo dedicado a recuperación de información esta medida es válida para comparar documentos. Sin embargo, no es demasiado adecuada para comparar consultas con documentos debido a las diferencias de tamaño; en este último caso se hace necesario "escalar" de algún modo los valores de P a fin de permitir una comparación razonable con Π.

[2] Es decir, $PiRo(d_1, d_2) = PiRo(d_2, d_1)$.

[3] `http://www.cs.cityu.edu.hk/~cssamk/gencomp/GenCompress1.htm`

[4] `http://www.lifl.fr/~varre/TD/tdsoft.html`

Q sea mínima. En su trabajo describen un método eficiente para obtener una solución aproximada para dicha búsqueda.

En resumen, *blindLight* es una técnica que utiliza vectores de pesos asignados a *n*-gramas de caracteres para representar documentos sin requerir ningún *corpus* para determinar dichos pesos. Para ello, emplea una medida de la significatividad de los *n*-gramas que puede ser calculada a partir de su frecuencia relativa de uso en un único documento. Los vectores así obtenidos pueden ser comparados no recurriendo a operaciones vectoriales sino empleando un nuevo vector "híbrido" resultado de la intersección de los vectores a comparar. De este modo, al relacionar la significatividad total de este nuevo vector con la de los vectores originales es posible obtener dos medidas asimétricas combinables linealmente para construir nuevas funciones de asociación entre documentos o entre documentos y consultas.

Como se verá en los siguientes capítulos el hecho de que la información para describir un documento se obtenga únicamente del propio documento sin recurrir a lo colección que lo incluye ni a un *corpus* externo no resulta en perjuicio de los resultados sino que estos son comparables[1] a los de otras técnicas que sí utilizan información de la colección para construir los vectores documentales.

3 Diferencias entre *blindLight* y otras técnicas PLN

Así pues, *blindLight* permite obtener para documentos escritos en cualquier lenguaje natural un "genoma" representado mediante un vector de *n*-gramas de caracteres. Cada *n*-grama (cada "gen") tiene asociado un peso que indica su significatividad en el texto original, significatividad calculada empleando un estadístico generalizado a *n*-gramas (Ferreira da Silva y Pereira Lopes 1999).

En este sentido *blindLight* diferiría muy poco de diversas implementaciones del modelo vectorial que han empleado como términos *n*-gramas de caracteres extraídos de los documentos. No obstante, a diferencia de tales implementaciones, la nueva técnica aquí descrita no emplea medidas de similitud "tradicionales" como por ejemplo la función del coseno. En su lugar, plantea la posibilidad de "hibridar" los vectores correspondientes a dos documentos obteniendo un tercer vector "artificial" y, mediante el mismo, obtener dos medidas asimétricas, Π y P, susceptibles de ser combinadas para constituir distintas medidas de similitud documental.

De este modo, resulta muy simple aplicar *blindLight* (al igual que el modelo vectorial) a tareas de clasificación, categorización o recuperación de información pero, además de éstas, también es posible emplear el "genoma" extraído de un documento para obtener resúmenes automáticos.

En cuanto método de PLN que emplea *n*-gramas de caracteres pueden encontrarse ciertas similitudes entre *blindLight* y técnicas como *Acquaintance* (Damashek 1995), *Highlights* (Cohen 1995) o la descrita por Neto *et al.* (2000). Sin embargo, tales similitudes son superficiales. Así, en el caso de *Acquaintance* se utilizan como pesos las frecuencias relativas de los *n*-gramas y no es aplicable a la obtención de resúmenes automáticos. *Highlights*, sólo permite obtener términos clave (no resúmenes) y se basa en la utilización de un "contexto" para cada documento al contrario que *blindLight* que extrae resúmenes empleando únicamente el documento a resumir. Por lo que respecta a la propuesta de Neto *et al.*

[1] A excepción, de momento, del sistema de recuperación de información como se verá en el capítulo 6.

únicamente obtiene resúmenes automáticos y lo hace siguiendo un enfoque análogo a la recuperación de información extendido a la recuperación de sentencias.

En cuanto a las propuestas de Eli Abir (2003a, 2003b, 2003c y 2004) (Abir *et al.* 2002) existe una aparente semejanza debido a la metáfora del "genoma documental" y la utilización de *n*-gramas de caracteres. No obstante, Abir no utiliza la misma técnica de ponderación que se emplea en *blindLight*, no parece aprovechar la posibilidad de "combinar" los genomas de distintos textos y no está claro si sus técnicas permiten obtener resúmenes automáticos ni en qué modo lo harían.

En definitiva, *blindLight*, al igual que otras técnicas, utiliza vectores de *n*-gramas de caracteres para representar documentos. Sin embargo, a diferencia de ellas emplea estos estadísticos generalizados para obtener los pesos de tales *n*-gramas y parece ser la única propuesta que plantea calcular un vector único para cada documento con independencia de la colección en la que se incluya. Además, no utiliza ninguna de las medidas de similitud habituales en los modelos vectoriales sino una basada en el uso de vectores obtenidos por combinación de otros y que en modo alguno son centroides. Por otro lado, el "genoma documental" va más lejos que la simple metáfora al poder combinarse con el texto original para producir resúmenes automáticos.

4 Semántica subyacente a los vectores *blindLight*

Antes de continuar es necesario dar soporte a la siguiente afirmación formulada por el autor en su tesis y sobre la que se apoyan todas las aplicaciones posteriores:

> …*Puesto que tales vectores almacenan ciertos aspectos de la semántica subyacente a los textos originales, el mayor o menor grado de similitud entre los mismos constituye un indicador de su nivel de relación conceptual*…

Para ello se debe aclarar en primer lugar el modo en que se utiliza aquí el término "semántica" y la forma más sencilla de hacerlo es señalando qué es lo que NO se afirma acerca de la nueva técnica propuesta por el autor:

- Al hablar de semántica no se pretende establecer ningún tipo de relación con la rama de la lingüística del mismo nombre. Ya se dijo con anterioridad que *blindLight* pretende, por el contrario, ser una técnica válida para escenarios en los que el conocimiento acerca de un idioma sea nulo.
- Tampoco se trata de establecer paralelismos con problemas de PLN tales como desambigüación o construcción automática de tesauros. Es necesario insistir en que la técnica no tiene como objetivo extraer ningún tipo de conocimiento estructurado a partir de los textos procesados.
- De igual modo *blindLight* tampoco tiene como objetivo facilitar la construcción automática de ontologías ni comparte rasgo alguno con la Web Semántica.

Así pues, ¿en qué sentido emplea el autor el término "semántica"? De un modo semejante al que lo hacen Susan Dumais *et al.* (1988, p. 282):

> *Damos por supuesto que bajo los datos de uso de palabras existe algún tipo de estructura semántica "latente" parcialmente oculta por la variabilidad en la elección de esas palabras. Utilizamos técnicas estadísticas para estimar dicha estructura latente y eliminar el "ruido".*

De manera similar, la presunción básica de este trabajo es que al descomponer el texto en *n*-gramas de caracteres y estimar una medida de la significatividad de dichos

n-gramas en un documento se dispone de una información no sólo valiosa sino fácilmente comparable con la de otros documentos a fin de determinar el grado de similitud a un nivel que va más allá de las simples palabras y que puede calificarse de "conceptual" o "semántico".

Como apoyo parcial a esto hay que señalar que se han desarrollado distintas iniciativas basadas en el uso de *n*-gramas para llevar a cabo tareas en las que habitualmente se requeriría el juicio de un humano. Así, Kishore Papineni *et al.* (2002) han desarrollado un método, *BLEU*, basado en *n*-gramas de palabras para evaluar la calidad de traducciones automáticas comparándolas con distintas traducciones de referencia producidas por humanos. Por otro lado, Chin-Yew Lin y Eduard Hovy (2003) demuestran que la utilización de *n*-gramas[1], nuevamente de palabras, permite evaluar con gran precisión la calidad de un resumen automático. Este trabajo evolucionó posteriormente hasta la implementación de una herramienta para realizar dicha evaluación: *ROUGE* (Lin 2004a).

En lo que resta de este apartado se va a describir un pequeño experimento que trata de sustentar la afirmación con que se comenzaba. Dicho experimento ha involucrado el uso de *blindLight* para clasificación de documentos. Los detalles sobre la implementación de dicha técnica se darán en capítulos posteriores y allí se proporcionarán más resultados. Los que se muestran aquí pretenden únicamente resultar suficientemente elocuentes acerca de la capacidad de *blindLight* para "modelar" la semántica subyacente a un texto.

4.1 Clasificación automática de (mini)corpora paralelos

Una de las utilidades más inmediatas de *blindLight* es la clasificación automática de documentos para lo cual se utiliza la medida de asociación *PiRo* mostrada en la página 66 (véase ecuación 12) y un algoritmo similar en ciertos aspectos al propuesto por R.A. Jarvis y Edward Patrick (1973). Si *blindLight* conservase realmente la semántica de los documentos entonces las clasificaciones de *corpora* paralelos deberían ser muy similares con independencia del idioma empleado en cada *corpus* y, además, plausibles según criterios de clasificación humanos.

En el experimento que se describe a continuación se puso a prueba semejante hipótesis. Para ello se buscaron en primer lugar documentos disponibles en formato electrónico y para los cuales existiesen traducciones en, al menos, los siguientes idiomas[2]: castellano (`ES`), finés (`FI`), francés (`FR`), hebreo (`HE`), holandés (`NL`), inglés (`EN`) y japonés (`JA`). Los documentos finalmente seleccionados fueron los siete siguientes:

- `CreativeCommons`: Licencia *Creative Commons* en su versión Atribución-No Comercial-Compartir en Igualdad.
- `Genesis`: Génesis 1:1-3:24.
- `GNU-GPL`: Licencia *GPL*.

[1] En particular unigramas y bigramas.

[2] Esos siete idiomas cubren las siguientes familias lingüísticas: indoeuropea (castellano, francés, holandés e inglés), urálica (finés), afroasiática (hebreo) y japonesa (japonés). Por lo que respecta a los idiomas indoeuropeos, el francés y el castellano son lenguas romances mientras que el holandés y el inglés son germánicas. *A priori*, sería de esperar encontrar similitudes entre las clasificaciones obtenidas para francés y castellano o para holandés e inglés debido a la mayor relación existente entre dichos lenguajes. En caso de encontrarse similitudes entre clasificaciones de documentos procedentes de familias distintas podría apoyarse con cierta confianza la capacidad de *blindLight* para conservar rasgos semánticos subyacentes al texto e independientes de características de un idioma o familia de idiomas en particular.

- `GoogleTermsOfService`: Condiciones de Servicio de *Google* en su versión de septiembre de 2004.
- `MSNTermsOfService`: Condiciones de Uso de *MSN* en su versión de abril de 2003.
- `UN-ConventionSaleGoods`: Convención de Viena sobre Compraventa Internacional de Mercaderías.
- `UN-HumanRights`: Declaración Universal de los Derechos Humanos.

Los documentos originales en inglés tienen longitudes bastante diferentes variando entre las 1.662 palabras de `GoogleTermsOfService` a las 11.182 de `MSNTermsOfService`. A fin de evitar que el tamaño pudiese afectar a los resultados algunos documentos fueron truncados para aproximar su longitud (en bytes) a la de aquellos más pequeños. De este modo el *corpus* final en inglés tuvo las siguientes características:

- `CreativeCommons`: Truncado a partir del apartado 5, tamaño final: 1.723 palabras.
- `Genesis`: Completo, tamaño final: 2.204 palabras.
- `GNU-GPL`: Truncado a partir del segundo párrafo del apartado 7, tamaño final: 1.855 palabras.
- `GoogleTermsOfService`: Completo, tamaño final: 1.662 palabras.
- `MSNTermsOfService`: Truncado a partir de la lista de acciones no permitidas, tamaño final: 1.670 palabras.
- `UN-ConventionSaleGoods`: Truncado a partir del artículo 20, tamaño final: 1.816 palabras.
- `UN-HumanRights`: Completo, tamaño final: 1.746 palabras.

Para construir el resto de *corpora* se utilizaron como "puntos de corte" los mismos que se usaron en los documentos en inglés a fin de garantizar que los textos finales fuesen realmente paralelos. Por otro lado, a fin de aplicar la técnica *blindLight* del mismo modo en todos los idiomas los textos en japonés y hebreo fueron transliterados[1]. Dado que el transliterador de Kanji a Romaji utilizado[2] producía la salida íntegramente en minúsculas todos los textos del resto de idiomas fueron también convertidos a minúsculas.

Así pues, una vez truncados los documentos necesarios en cada idioma, transliterados los documentos en hebreo y japonés y convertidos todos los textos a minúsculas se disponía de siete *corpora* paralelos garantizando además que todos los documentos tenían el mismo tamaño aproximado en un idioma "patrón". De este modo, las posibles diferencias de longitud en el resto de idiomas deberían ser mínimas y sólo podrían ser atribuíbles a efectos de la traducción.

[1] En teoría empleando *Unicode* se podría aplicar la técnica de manera directa sobre los textos en hebreo y japonés. No obstante, dada la distinta naturaleza de los sistemas de escritura se optó por la transliteración (esto es, su representación empleando el alfabeto latino) a fin de usar *n*-gramas del mismo tamaño en todos los idiomas.

[2] `http://www.j-talk.com/nihongo/`

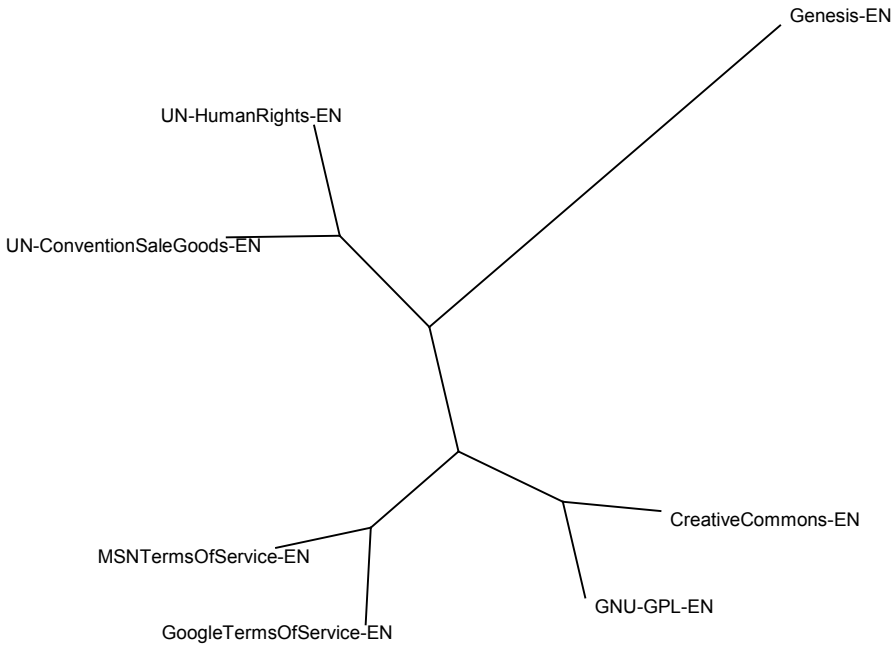


Fig. 36 Clasificación *blindLight* del corpus de documentos escritos en inglés.

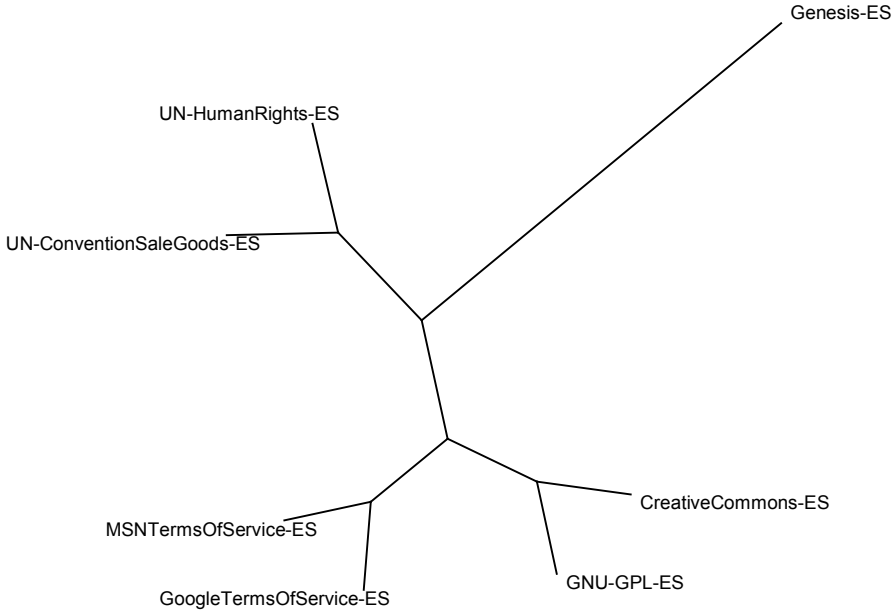


Fig. 37 Clasificación *blindLight* del corpus de documentos escritos en castellano.

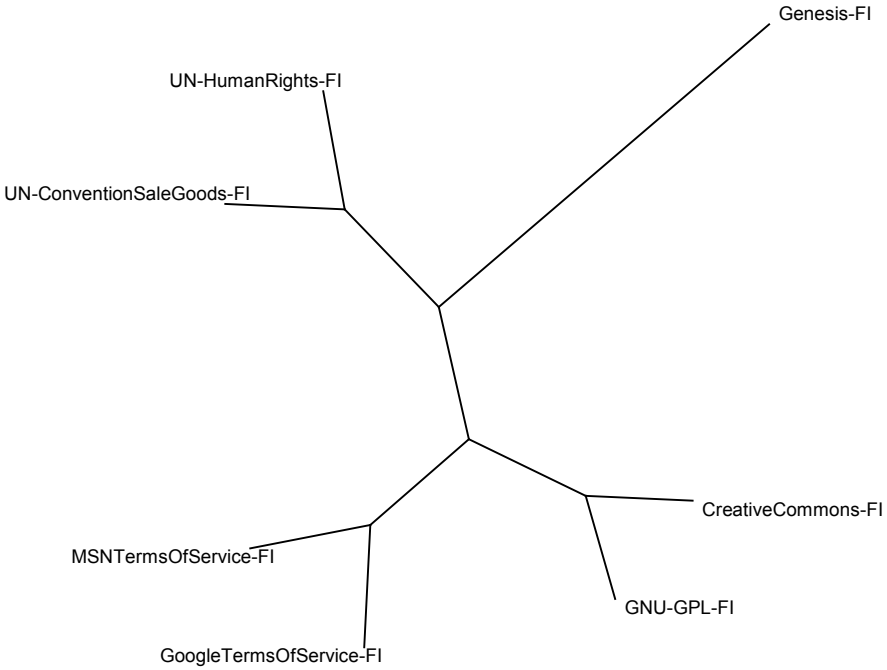


Fig. 38 Clasificación *blindLight* del corpus de documentos escritos en finés.

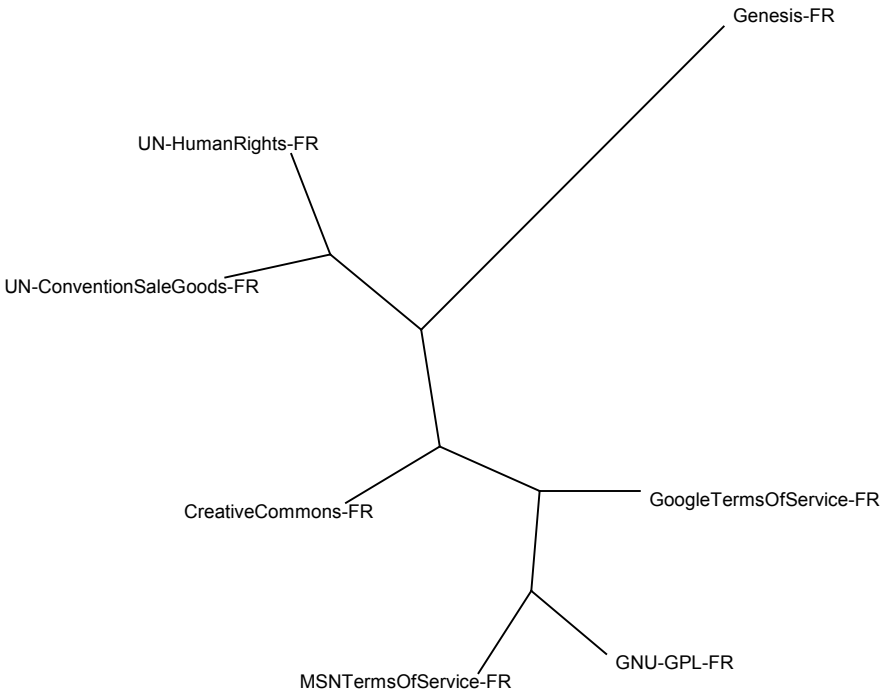


Fig. 39 Clasificación *blindLight* del corpus de documentos escritos en francés.

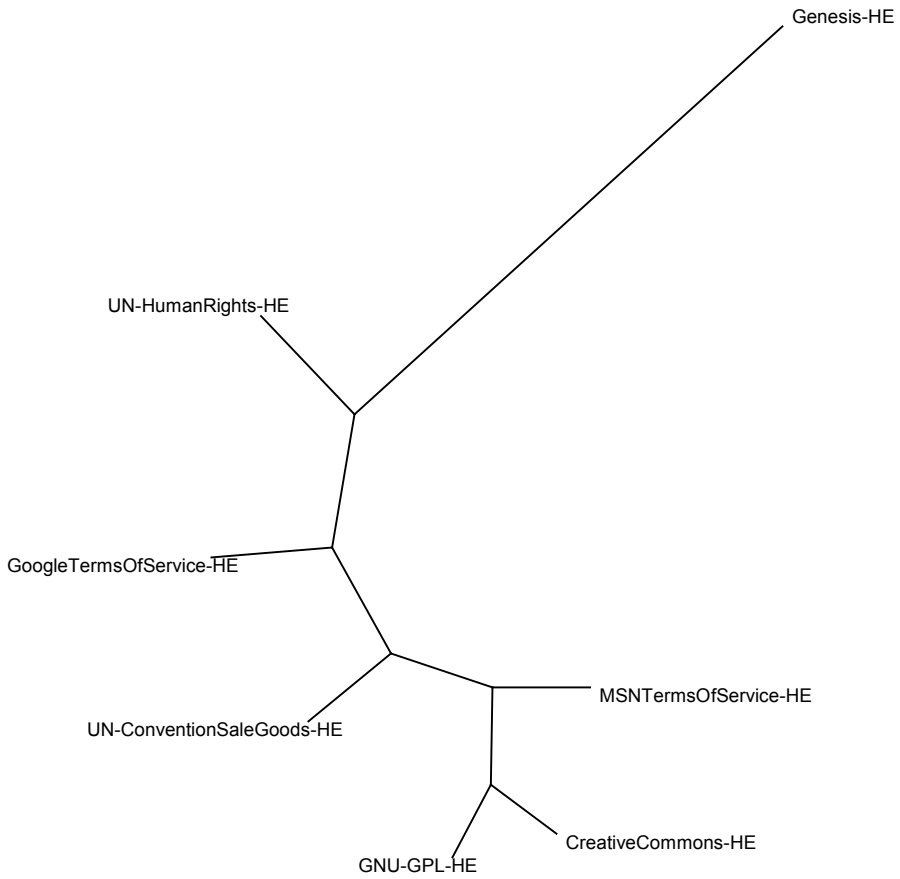


Fig. 40 Clasificación *blindLight* del corpus de documentos escritos en hebreo.

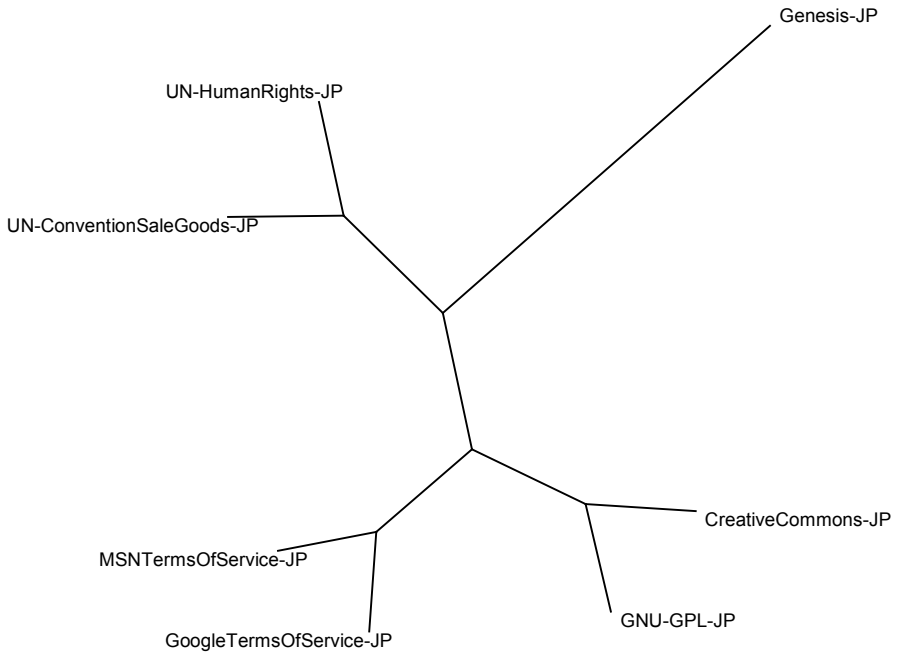


Fig. 41 Clasificación *blindLight* del corpus de documentos escritos en japonés.

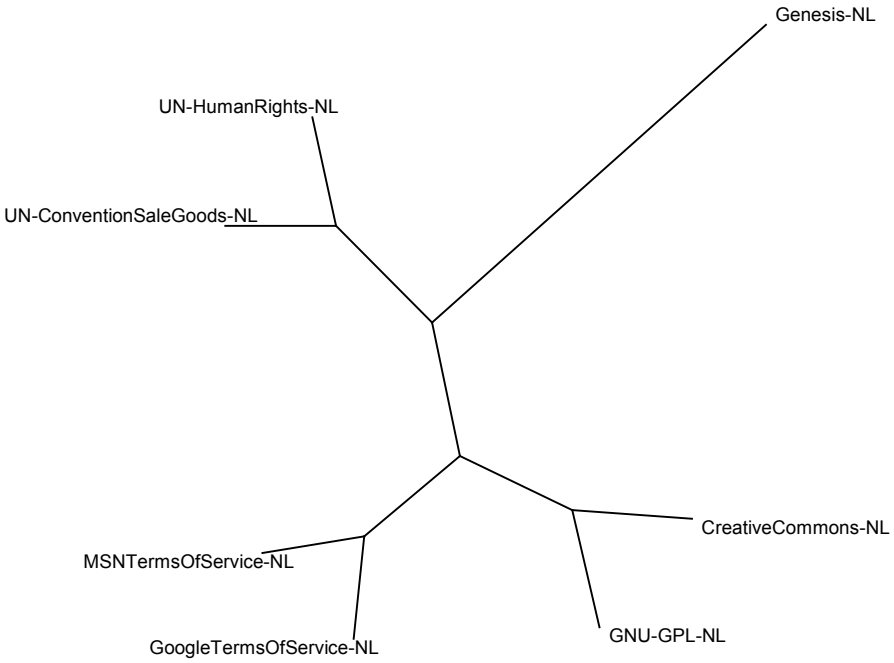
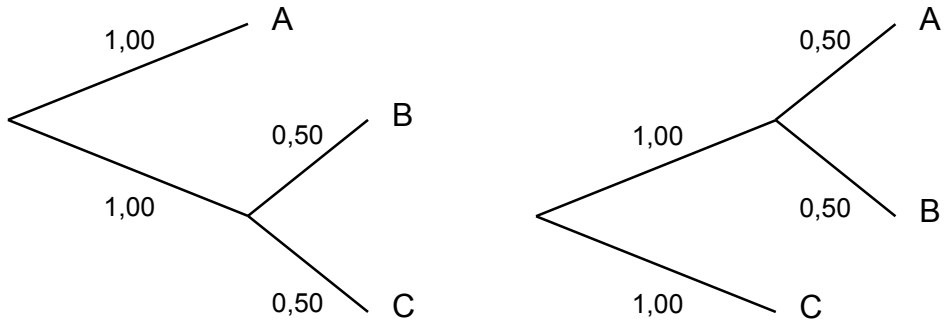


Fig. 42 Clasificación *blindLight* del corpus de documentos escritos en holandés.

Como se puede observar en las figuras Fig. 36 a Fig. 42 que muestran las distintas clasificaciones éstas son topológicamente equivalentes para todos los *corpora* a excepción del francés y del hebreo. Sin embargo, un simple parecido no es suficiente y es necesario evaluar numéricamente la similitud (o su ausencia) entre los distintos árboles obtenidos. Para ello se han empleado dos medidas de comparación entre árboles, la primera es la distancia *branch score* descrita por Mary Kuhner y Joseph Felsenstein (1994, p. 461) y la segunda, propuesta por el autor, está basada en el coeficiente de correlación de Spearman.



Árbol 1 **Árbol 2**

Fig. 43 La distancia branch score entre estos árboles es 0,5 y el coeficiente de correlación -0,25.

A fin de explicar el cálculo de ambas medidas se emplearán los árboles mostrados en la Fig. 43. Para el cálculo de la distancia *branch score* hay que determinar, en primer lugar, todas las posibles particiones que se pueden establecer en cada árbol teniendo en cuenta que no existe ningún orden pre-establecido entre las distintas ramas.

Así, para los dos árboles del ejemplo es posible obtener las particiones `{A|B,C}`, `{A,B|C}` y `{A,C|B}` (véase Fig. 44). Naturalmente habrá casos más complejos donde no

todas las particiones posibles en un árbol sean posibles en el otro. Seguidamente, se determina la longitud de la rama de cada partición teniendo en cuenta que si una partición no existiese en un árbol se le asignaría longitud cero. En este caso particular las particiones tendrían las longitudes que se muestran en Fig. 45.

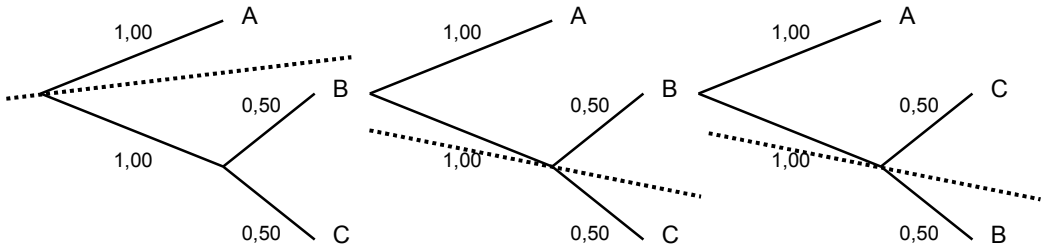


Fig. 44 Particiones del primer árbol {A | B,C}, {A,B | C} y {A,C | B} con distancias 1.00, 0.50 y 0.50 respectivamente.

Posteriormente, se calcula el cuadrado de las diferencias de las longitudes de cada partición y se suman. Así, en este ejemplo la distancia *branch score* sería de 0,50. Esta distancia es 0,00 para árboles idénticos y aumenta a medida que los árboles difieren entre sí. Sin embargo, depende del tamaño de los árboles comparados de tal manera que no es posible comparar distancias para parejas de árboles diferentes (Kuhner y Felsenstein, p. 461). Por esa razón el autor ha optado por utilizar una *branch score* "normalizada" al dividir la raíz cuadrada de la puntuación obtenida entre la suma de las longitudes de todas las ramas de los dos árboles comparados. En el ejemplo anterior la puntuación normalizada sería de 0,2357.

	Árbol 1	Árbol 2
{A | B, C}	1,00	0,50
{A, B | C}	0,50	1,00
{A, C | B}	0,50	0,50

Fig. 45 Longitudes de las particiones en cada árbol.

La medida propuesta por el autor, por su parte, se basa en el coeficiente de Spearman de correlación entre listas ordenadas de elementos. La idea es simple, dados dos árboles que contienen los mismos nodos es posible definir en ambos dos listas idénticas de posibles parejas. En el ejemplo utilizado hasta el momento la lista sería `{AB,AC,BC}`. Para cada árbol se puede obtener la distancia que hay que "recorrer" para ir de un elemento de cada pareja al otro (véase Fig. 46).

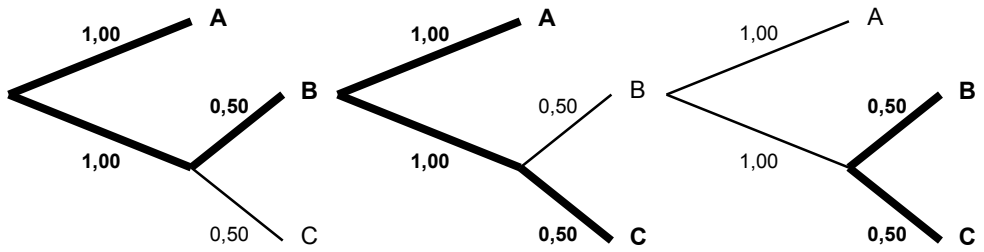


Fig. 46 Distancias "sobre" el árbol entre A y B (2.50), A y C (2.50) y B y C (1.00).

Una vez obtenidas las distancias para cada pareja se asigna a cada una un *ranking* en cada árbol y se procede a la determinación del correspondiente coeficiente de Spearman (véase Fig. 47). En el caso del ejemplo el coeficiente obtenido es -0,125 lo cual significa que apenas hay correlación entre ambos árboles.

	Distancia "sobre" A1	*Ranking* A1	Distancia "sobre" A2	*Ranking* A2	Diferencia *rankings*	Cuadrado diferencias
AB	2,50	2,5	1,00	1	1,5	2,25
AC	2,50	2,5	2,50	2,5	0	0
BC	1,00	1	2,50	2,5	1,5	2,25
				Suma cuadrados diferencias		4,50
				Número de parejas (n)		3

Coeficiente de correlación $1-\frac{6\sum_{i=1}^{n}(r_i^1-r_i^2)^2}{n^3-n}=1-\frac{6\cdot 4{,}50}{3^3-3}=-0{,}125$

Fig. 47 Cálculo del coeficiente de correlación de Spearman para lo árboles 1 y 2. Obsérvese que, en caso de empate, el *ranking* asignado es la media de los *rankings* teóricos.

La principal ventaja de esta medida de comparación es que sus valores varían entre -1 y 1, significando estos valores extremos una correlación perfecta (negativa o positiva) y los valores próximos a cero la ausencia de correlación. Por otro lado, puesto que las distancias entre los elementos de la pareja se miden "sobre" el árbol no se pierde totalmente la información topológica.

Las similitudes entre las clasificaciones para los distintos *corpora* determinadas mediante ambas medidas son las que se muestran en Fig. 48. Como se puede ver, existen varias parejas de clasificaciones en las que tanto la medida *branch score* como el coeficiente de Spearman coinciden en señalar una enorme similitud (p.ej. las formadas por inglés, español, finés y holandés). No obstante, también existen otras en las que sólo una de las medidas señala una similitud más o menos importante; sin embargo, en tales casos siempre interviene alguno de los siguientes idiomas: francés, hebreo o japonés.

	BRANCH SCORE		SPEARMAN		
	Branch Score	*Branch Score* normalizada	Coef. Correlación	Fiabilidad	Correlación
EN-ES	0,26	0,015	0,98	99%	Casi perfecta
EN-FI	0,43	0,019	0,97	99%	Casi perfecta
EN-FR	6,63	0,076	0,93	99%	Muy fuerte
EN-HE	114,44	0,184	0,75	99%	Fuerte
EN-JP	2,63	0,053	0,71	99%	Fuerte
EN-NL	0,19	0,012	0,97	99%	Casi perfecta
ES-FI	0,50	0,022	0,90	99%	Muy fuerte
ES-FR	5,65	0,073	0,90	99%	Muy fuerte
ES-HE	116,67	0,191	0,77	99%	Fuerte
ES-JP	1,60	0,044	0,75	99%	Fuerte
ES-NL	0,40	0,019	1,00	99%	Perfecta
FI-FR	6,33	0,076	0,92	99%	Muy fuerte
FI-HE	116,27	0,188	0,68	99%	Fuerte
FI-JP	1,44	0,041	0,61	99%	Fuerte
FI-NL	0,21	0,014	0,90	99%	Muy fuerte
FR-HE	122,14	0,194	0,69	99%	Fuerte
FR-JP	6,27	0,086	0,61	99%	Fuerte
FR-NL	6,58	0,076	0,91	99%	Muy fuerte
HE-JP	125,29	0,210	0,41	Rechazar	Inexistente
HE-NL	115,56	0,185	0,76	99%	Fuerte
JP-NL	2,04	0,047	0,75	99%	Fuerte

Fig. 48 Medidas de la similitud entre las clasificaciones obtenidas para los distintos *corpora*.

Se muestran con sombreado gris aquellas clasificaciones con mayor similitud de acuerdo a la distancia *branch score* normalizada y en negrita aquellas con mayor grado de similitud de acuerdo al coeficiente de correlación de Spearman.

Así, los resultados obtenidos con los documentos en japonés son muy similares a los obtenidos con el resto de idiomas de acuerdo con la medida *branch score* y no tan parecidos aplicando Spearman. En cambio, los resultados para el francés no son topológicamente parecidos (*branch score*) pero sí resultan muy similares mediante Spearman. Y por lo que respecta al hebreo, es el idioma cuyos resultados parecen estar menos relacionados con los obtenidos para otras lenguas.

En el caso del hebreo y japonés la principal razón para la menor correlación (a pesar de una obvia semejanza topológica en el caso del japonés) es, muy probablemente, el uso por parte de ambos de un sistema de escritura muy diferente al empleado por el resto de idiomas comparados obligando a una transliteración de los textos.

Así, el hebreo utiliza un alfabeto de 22 letras. Salvo en unos pocos casos especiales[1] puede decirse que el alfabeto permite representar tan sólo consonantes. Para indicar los sonidos vocálicos se utiliza un sistema de marcas (fundamentalmente puntos) situados en las proximidades de cada letra. Este sistema, conocido como *niqqud* o *nikkud*, también permite señalar qué sílaba de la palabra lleva el acento. La utilización del *niqqud* es opcional y la mayor parte de los textos escritos en hebreo, en particular los utilizados en este experimento, utilizan únicamente consonantes. Este hecho puede tener una gran influencia al comparar los resultados obtenidos para el hebreo con los de otras lenguas. Mientras que los textos empleados para el resto de idiomas recogen mucha información acerca de la vocalización de cada texto[2] en el caso de los documentos escritos en hebreo esa información simplemente no existe.

Por otro lado, dos de los documentos utilizados no se correspondían con la fidelidad necesaria al resto de traducciones utilizadas. El primer caso es el de la versión hebrea de la *"Convención de las Naciones Unidas sobre los contratos de compraventa internacional de mercaderías"* para la que no se pudo localizar ninguna versión del documento que incluyese el preámbulo (115 palabras en la versión inglesa). El segundo es el documento con las condiciones de uso de *Google* en su versión para Israel. Los dos últimos párrafos del apartado *"Renuncia a garantías"* (תעודת אחריות) no existen. El apartado de *"Limitación de responsabilidad"* (האחריות מגבלו) está vacío. En el apartado *"Solicitud de eliminación de vínculos o materiales en caché"* (בקשה להסרת קישורים או חומרי מאוחסנים) tan sólo aparece el primer párrafo pero no se citan los principios que emplea *Google* para decidir sobre las solicitudes de eliminación. El apartado *"Condiciones varias"* (תנאים שונים) también está vacío. En definitiva, 899 palabras de las 1662 del documento original en inglés no aparecen en la traducción.

En resumen, la versión hebrea del documento de la Convención de Viena sobre compraventa de mercaderías carece de, aproximadamente, un 6% del texto original y el correspondiente a las Condiciones de Servicio de Google de un 54%. Este hecho, unido a la ausencia de información vocálica, sin duda tiene un impacto en la taxonomía finalmente obtenida y en su falta de concordancia con el resto de idiomas estudiados.

Por lo que se refiere a los documentos escritos en japonés, hasta donde ha podido comprobar el autor, las traducciones son razonablemente fieles. Por otro lado, hay que

[1] Ciertas letras son mudas o vocales dependiendo de su combinación con sonidos vocálicos o al aparecer al final de palabra.

[2] El español escrito, por ejemplo, recoge de forma casi total la vocalización y acentuación de las palabras. Otros idiomas recogen mucha información "fonética" pero no toda la necesaria. Pensemos en el inglés con sus *bough* /baʊ/, *bought* /bɔːt/, *cough* /kɒf/, *though* /ðəʊ/ o *through* /θruː/.

señalar que ciertas características del japonés escrito y de su pronunciación resultan difíciles de manejar en su transliteración y, por tanto, han podido influir en los resultados finales.

En primer lugar, el japonés escrito puede emplear hasta cuatro tipos distintos de caracteres: *kanji*, *hiragana*, *katakana* y *rōmaji*. Los primeros son caracteres chinos adaptados a la escritura de sustantivos, adjetivos, verbos y nombres propios japoneses. *Hiragana* y *katakana* son silabarios utilizándose el segundo en particular para transcribir palabras y nombres no japoneses. Por último, el *rōmaji* emplea caracteres latinos y uno de sus posibles usos es la transliteración de los anteriores sistemas de escritura.

La transliteración de los silabarios *hiragana* y *katakana* no supone mayor problema. Sin embargo, la del *kanji* resulta más compleja puesto que un único carácter puede tener varias "lecturas" en función de la palabra de la que forma parte y, por tanto, del contexto (véase Fig. 49). El transliterador utilizado emplea el léxico del proyecto *EDICT*[1] de Jim Breen para segmentar el texto y proporciona, por tanto, una única romanización que, presumiblemente, es la más adecuada para cada contexto. Sin embargo, el autor no es en absoluto un experto en japonés por lo que no se puede asegurar que los documentos transliterados de manera automática estén totalmente libres de errores que podrían haber influido en los resultados finales.

米	Hiragana Katakana	Pronunciación	Tipo pronunciación
	ベイ	bei	on'yomi
	マイ	mai	
	メエトル	meetoru	
	こめ	kome	kun'yomi
	よね	yone	
	は	ha	nanori
	べ	be	
	まべ	mabe	
	め	me	
	よ	yo	
	よな	yona	
	よの	yono	
	よま	yoma	

国	Hiragana Katakana	Pronunciación	Tipo pronunciación
	コク	koku	on'yomi
	くに	kuni	kun'yomi
	くな	kuna	nanori
	こ	ko	

Fig. 49 Distintas lecturas de los caracteres de la "palabra" japonesa 米国 (べいこく, *beikoku*, EEUU).

Los caracteres japoneses poseen tres tipos distintos de pronunciaciones: *on'yomi* (pronunciación china), *kun'yomi* (pronunciación japonesa) y *nanori*. La primera es la pronunciación aproximada del carácter chino original. *Kun'yomi* es la pronunciación de una palabra japonesa equivalente al carácter chino importado. *Nanori* es la forma en que puede pronunciarse el *kanji* cuando se utiliza dentro de un nombre propio. Un transliterador de japonés debe determinar en qué contexto se está utilizando un carácter a fin de proporcionar su pronunciación correcta; esto habitualmente se hace empleando diccionarios que recogen la versión *kanji* y el correspondiente *hiragana/katakana* para cada entrada.

[1] http://www.csse.monash.edu.au/~jwb/j_edict.html

Por lo que respecta al francés se ha realizado un análisis *post mortem* de los documentos a fin de encontrar causas para las discrepancias, en particular con los idiomas indoeuropeos. Tras ese análisis se ha encontrado lo siguiente:

- Condiciones de Uso de *Google*: Unas pocas partes del texto original en inglés no se incluyen en la versión francesa, no obstante se trata de una traducción razonablemente fiel.
- Creative Commons: los puntos e) y f) del apartado 4 (*Restricciones*) no aparecen en la versión francesa de la licencia (232 palabras en el original en inglés). El apartado 5 (*Renuncia de responsabilidades*) presenta una redacción totalmente diferente (83 palabras en el original). De este modo, un 13% del texto original no aparece en la versión en francés y un 5% ha sido redactado de forma totalmente diferente adquiriendo un mayor peso dentro del documento traducido (11%). En resumen, la licencia Creative Commons en francés no es una traducción fiel de la licencia en inglés sino una adaptación.
- El resto de documentos (Condiciones de Uso de *MSN*, *GNU General Public License*, Génesis, Convención de Viena y Declaración Universal de Derechos Humanos) son traducciones fieles.

Las diferencias entre la clasificación de los documentos franceses y las obtenidas para el resto de idiomas tal vez puedan atribuirse a las diferencias perceptibles que se han descrito entre la licencia *CCPL* en su versión francesa.

No obstante, a pesar de lo señalado para los documentos en francés, hebreo y japonés, puede concluirse que al clasificar un conjunto de *corpora* paralelos utilizando *blindLight* se obtienen de manera sistemática clasificaciones idénticas o muy similares con independencia de la familia lingüística y la longitud de los documentos[1]. Dichas clasificaciones, además, son plausibles según criterios humanos puesto que tienden a agrupar documentos de contenido semejante y mantener separados textos de temática muy distinta en todos los idiomas.

El autor considera que tales resultados permiten sustentar lo que se afirmaba al comienzo del apartado, a saber, que al aplicar la técnica aquí descrita sobre texto natural se obtienen vectores que conservan ciertos aspectos de la semántica latente en los documentos originales permitiendo una comparación a un nivel conceptual.

En los siguientes capítulos se presentará la aplicación de la técnica a la clasificación y categorización de documentos, a la recuperación de información y a la obtención de resúmenes automáticos concluyendo de ese modo la demostración de la tesis del autor.

[1] Tan sólo en inglés los documentos tienen aproximadamente la misma longitud en *bytes*, en el resto de idiomas los documentos difieren en tamaño.

CAPÍTULO 4

CLASIFICACIÓN DE DOCUMENTOS CON *BLINDLIGHT*

Un primer mecanismo para enfrentarse a una colección muy grande de documentos es su clasificación, es decir, su división en grupos de documentos más pequeños y homogéneos que permitan deducir la estructura subyacente a la colección y faciliten su exploración. El problema de la clasificación no supervisada de colecciones de documentos no es nuevo y los beneficios que aporta al campo de la recuperación de información son bien conocidos. Existe una gran variedad de métodos de clasificación y se dispone de técnicas que permiten evaluar la calidad de las clasificaciones obtenidas, ya sea estudiando características internas de las mismas o comparándolas con clasificaciones "externas". En este capítulo se repasarán brevemente algunas de las principales técnicas de clasificación automática de documentos para, seguidamente, presentar la forma en que es posible aplicar la técnica propuesta por el autor a este problema. A continuación, se describirá una serie de experimentos cuyos resultados serán comparados con los obtenidos con otras técnicas "convencionales" demostrando que, efectivamente, blindLight es capaz de obtener resultados semejantes e incluso mejores que los alcanzados por métodos ad hoc.

1 El problema de la clasificación

El problema de clasificar de manera no supervisada un conjunto de patrones, *a priori* grande, en un número reducido de grupos (*clusters*) que exhiban características similares es conocido como *clustering*, clasificación no supervisada o agrupamiento. Este problema se manifiesta en multitud de campos, incluyendo la recuperación de información[1], y admite varias aproximaciones. Sin embargo, no es objetivo de este trabajo analizar de manera

[1] La aplicación de técnicas de clasificación a la recuperación de información se basa en la denominada *cluster hypothesis* expuesta originalmente por Jardine y van Rijsbergen (1971, p. 219): *"es intuitivamente plausible que la asociación entre documentos proporciona información acerca de la relevancia de los documentos respecto a las peticiones"* y posteriormente reformulada del siguiente modo: *"documentos estrechamente asociados tienden a ser relevantes para las mismas peticiones"* (van Rijsbergen 1979).

exhaustiva las distintas alternativas para llevar a cabo clasificación de documentos puesto que existe amplia bibliografía sobre el tema. Para adquirir una visión amplia del campo son muy recomendables tanto el tercer capítulo de *"Information Retrieval"* (van Rijsbergen 1979) como la revisión hecha por Jain, Murty y Flynn (1999). Los capítulos cuarto, y en menor medida el quinto y sexto, de *"Mining the Web"* (Chakrabarti 2003) están dedicados al problema de extraer información de colecciones de documentos hipertextuales.

No obstante, puesto que el autor afirma en su tesis que la técnica que propone facilita *"la clasificación […] de documentos […] con resultados similares a los de otros métodos […]"* antes de describir la forma en que es posible aplicar *blindLight* a este problema y los resultados obtenidos es necesario hacer un brevísimo repaso de las distintas técnicas disponibles para la clasificación de documentos así como de algunas colecciones habitualmente empleadas para probar tales métodos.

1.1 Clasificación de documentos

Un método de clasificación requiere, básicamente, (1) un modo de representar los documentos, (2) una medida de similitud entre dichas representaciones y (3) un algoritmo para construir los grupos de documentos basándose en la medida anterior.

La forma de representar los documentos para su clasificación y, de hecho, cualquier patrón es, habitualmente, *"un vector multidimensional donde cada dimensión corresponde a una única característica"* (Duda y Hart 1973, citado por Jain *et al.* 1999, p. 270). Como ya se dijo anteriormente, el modelo vectorial (ya sean los pesos binarios o reales) facilita un modo de representación de documentos muy conveniente para la implementación de métodos de clasificación.

En cuanto a las medidas de similitud ya se han mostrado varias (p.ej. los coeficientes de Dice o Jaccard o la función del coseno). Según Lerman (1970, citado por van Rijsbergen 1979, p. 30) muchas de estas medidas son monótonas entre sí[1]. Por tanto, aquellos métodos de clasificación que únicamente empleen el orden establecido entre los documentos, es decir, la mayor parte, obtendrán clasificaciones idénticas con independencia de la medida de similitud empleada.

Así, de los tres pasos necesarios para clasificar documentos (representación, cálculo de similitudes y construcción de los grupos) este apartado se centrará únicamente en el tercero revisando escuetamente algunas de las distintas aproximaciones posibles.

Jain *et al.* (1999) proporcionan una posible taxonomía de métodos de clasificación que, inicialmente, pueden diferenciarse en dos grandes grupos: los jerárquicos que producen un conjunto de particiones anidadas y los particionales que producen una única partición. En segundo lugar existen una serie de características "transversales" que afectan a ambos tipos. Así, por ejemplo, los algoritmos pueden ser (a) aglomerativos o divisivos[2] según comiencen con tantos grupos como documentos que van siendo "fusionados" o lo hagan con un único grupo que es dividido. (b) Exactos si un documento es asignado a un único grupo o borrosos (*fuzzy*) si existen grados de "pertenencia". (c) Deterministas si siempre se obtiene la misma clasificación para un conjunto de partida o estocásticos si se emplean métodos aleatorios para llegar al resultado final. Y (d) incrementales o no incrementales en

[1] Dos funciones son monótonas entre sí cuando al establecer una relación de orden sobre un conjunto de elementos ambas dan lugar a la misma ordenación.

[2] Ascendentes y descendentes según Chakrabarti (2003, p. 84)

función del número de documentos a clasificar[1]. Así pues, existe una flexibilidad enorme a la hora de especificar un algoritmo de clasificación y en consecuencia existen muchísimos métodos y sutiles variantes de los mismos. A continuación se presentarán algunos de los más conocidos.

Los métodos aglomerativos comienzan situando cada documento en su propio grupo y, de forma iterativa, van agrupando grupos en funcion de su similitud. Puesto que en cada iteración existe un número distinto de grupos y éstos están constituidos a su vez por grupos más pequeños este tipo de algoritmos son jerárquicos. Las dos principales variantes de este tipo de algoritmos son las denominadas *single-link* (Sneath y Sokal 1973) y *complete-link* (King 1967). En el primer caso la distancia entre dos grupos es el mínimo de las distancias entre todos los pares formados al emparejar un elemento del primer grupo con uno del segundo. En el caso del método *complete-link* esta distancia no es el mínimo sino el máximo. En ambos casos se agrupan iterativamente aquellos grupos que se encuentran a una distancia menor.

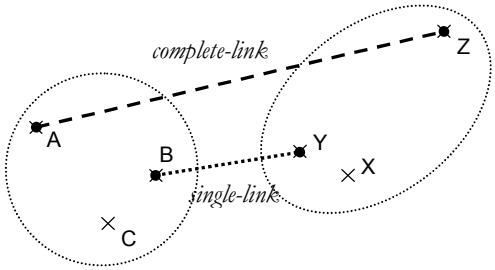


Fig. 50 Distancia entre grupos empleando *complete-link* y *single-link*.

Un segundo método muy conocido es el algoritmo de las *k*-medias que es particional, divisivo y, en su versión más simple, exacto (Duda y Hart 1973, citado por Jain *et al.* 1999). Este método clasifica una colección de patrones en *k* grupos donde el valor de *k* se establece *a priori*. Cuando se aplica a clasificación de documentos estos son representados como vectores y los grupos como el centroide de los documentos pertenecientes al mismo. En primer lugar deben establecerse *k* centroides de manera "arbitraria" asignándose cada documento de la colección al centroide más próximo. Una vez se han asignado todos los documentos se recalcula el centroide de cada grupo y se lleva a cabo una nueva fase de asignación. Este proceso se repite hasta que los cambios en los grupos obtenidos son mínimos. Una alternativa a este método es conocida como *k*-medoides donde no se utiliza el centroide del grupo sino aquel documento del mismo más próximo a éste.

Otro algoritmo interesante es el propuesto por Jarvis y Patrick (1973). Este método no sólo tiene en cuenta los documentos más próximos a uno dado sino también los vecinos que tienen en común, razón por la que también se denomina como método de "vecinos comunes" (*shared nearest neighbor clustering*). Este algoritmo requiere dos parámetros, *J* y *K*, donde *J* es el tamaño de la lista de vecinos para un punto dado y *K* es el número de vecinos comunes necesarios para formar un grupo. Según el método de Jarvis-Patrick dos documentos pertenecerán a un mismo grupo si ambos son vecinos y tienen, al menos, *K* vecinos en común.

---

[1] *"La clasificación incremental parte del supuesto de que es posible considerar los patrones de uno en uno y asignarlos a algún grupo ya disponible"* (Jain *et al.* 1999, p. 32) por lo que está especialmente indicada para colecciones muy grandes.

Además de métodos como los anteriores se han aplicado con éxito técnicas de computación flexible y probabilísticas. Por ejemplo, los Mapas Auto-Organizativos[1] (*Self-Organizing Maps* o *SOM*) se han utilizado en los proyectos *WEBSOM* (Honkela *et al.* 1996) y *SOMLib* (Rauber y Merkl 1999) y se han desarrollado métodos de clasificación jerárquicos y aglomerativos basados en la probabilidad condicionada de Bayes (Iwayama y Tokunaga 1995).

1.2 Evaluación de métodos de clasificación

Según van Rijsbergen (1979, p. 23) la clasificación en el campo de la recuperación de información se hace con un propósito y, por tanto, su bondad sólo puede medirse sobre la base del rendimiento en la fase de recuperación. De este modo, van Rijsbergen evita debatir acerca de las denominadas clasificaciones "naturales", aquellas similares a las que producirían de manera independiente distintos seres humanos.

No obstante, no es estrictamente necesario esperar a la fase de recuperación de información para evaluar un método de clasificación automática. De hecho, si se dispone de documentos que ya se han clasificado previamente (tal vez de manera manual) es posible calcular una serie de medidas para determinar la calidad de la clasificación automática: la **entropía**[2] y la **medida *F***[3] (Steinbach, Karypis y Kumar 2000) o la **pureza**[4] (Zhao y Karypis 2002). Por otro lado, en caso de no disponer de una clasificación previa puede calcularse la **similitud promedio**[5] (*overall similarity*) (Steinbach *et al.* 2000).

En cuanto a las colecciones de documentos que se utilizan con mayor frecuencia para evaluar nuevos métodos de clasificación podrían destacarse la colección de artículos de la revista *TIME*, las habituales *CACM*, *CISI*, *LISA* y *Cranfield*[6], diversas particiones

---

[1] Los Mapas Auto-Organizativos (Kohonen 1982) agrupan una serie de vectores de entrada sobre un "mapa", una red neuronal generalmente bidimensional o tridimensional, en el cual vectores "similares" aparecen en posiciones cercanas.

[2] La entropía es un criterio de evaluación externo puesto que depende de una clasificación previa con la que comparar la solución obtenida. Dado un grupo *j*, su entropía es E_j (véase ecuación al final de la nota) donde p_{ij} es la probabilidad de que un elemento de dicho grupo pertenezca a la clase *i*. Por su parte, la entropía del agrupamiento será la media ponderada de la entropía de todos los grupos (en función de la proporción entre el número de documentos del grupo y el total). $E_j = -\sum_i p_{ij} \cdot \log(p_{ij})$

[3] La medida *F* se utiliza habitualmente para evaluar sistemas de recuperación de información (véase pág. 139) y combina en un valor único las consabidas precisión y exhaustividad. En el caso de la evaluación de clasificaciones automáticas esta medida permite evaluar soluciones jerárquicas (al contrario que la entropía y la pureza que tan sólo sirven para soluciones sin grupos anidados). Para ello, se calculan los valores F_{ij} para cada grupo *j* y clase externa *i* entendiendo que la precisión es la fracción de documentos del grupo *j* que pertenecen a la clase *i* mientras la exhaustividad es la fracción de documentos de la clase *i* que aparecen en el grupo *j*. Posteriormente se determina el máximo valor *F* para cada clase y, por último, se calcula la media ponderada de estos valores.

[4] La pureza también permite evaluar una solución de agrupamiento por medio de una clasificación externa. No es más que la proporción entre el número de *ítems* pertenecientes a la clase dominante en un grupo y el tamaño de dicho grupo. Es decir, la pureza evalúa en qué medida un grupo de una clasificación automática contiene elementos de una única clase.

[5] La similitud promedio es una medida de evaluación interna y que, por tanto, no requiere ninguna clasificación con la que comparar la solución de agrupamiento. Se trata tan sólo de la similitud media entre cada par de documentos de un grupo.

[6] Se trata de colecciones para la evaluación de sistemas de recuperación de información pero que también se han usado para evaluar sistemas de clasificación automática puesto que los documentos están asignados a categorías predefinidas. La colección *CACM* consta de 3.204 artículos (título y resumen) publicados en

empleadas en las conferencias *TREC*[1], la colección de textos médicos *OHSUMED* (Hersh *et al.* 1994) y varias colecciones de artículos de la agencia *Reuters*[2].

2 Utilización de *blindLight* para la clasificación automática de documentos

La aplicación de *blindLight* a la clasificación automática de documentos es muy sencilla: los documentos se representan mediante vectores de *n*-gramas tal y como fueron descritos en el capítulo anterior mientras que la medida de similitud interdocumental es la denominada *PiRo* (véase ecuación 12 en página 66). Por lo que respecta a los métodos para obtener los grupos de documentos se han implementado dos algoritmos particionales, uno no incremental y otro incremental. En el primer caso se utiliza, además de la similitud interdocumental, información sobre los documentos "vecinos" de manera similar a la de Jarvis y Patrick (1973).

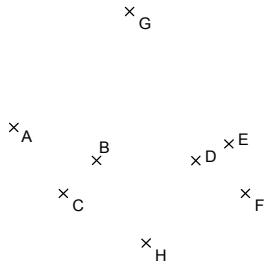


Fig. 51 Conjunto de puntos de ejemplo.

2.1 Algoritmo no incremental basado en *blindLight*

Una de las ideas más interesantes del algoritmo de clasificación de Jarvis y Patrick es la utilización de listas de vecinos "compartidos" para agrupar los patrones. En la implementación del método de clasificación *blindLight* no incremental se ha aplicado una idea similar pero sin recurrir a ningún tipo de lista de vecinos y utilizando en cambio toda la información disponible en la matriz de similitudes. Dicha matriz se construye en la fase inicial del algoritmo y almacena los valores *PiRo* para cada posible par de documentos[3]. De este modo, para cada documento D_i se dispone de un vector que contiene la similitud de D_i respecto al resto de documentos de la colección (véase Fig. 52). Estos vectores pueden asimilarse con el "comportamiento" de cada documento dentro de la colección y permitirían agrupar aquellos documentos que no sólo son similares entre sí sino que se "comportan" de manera similar respecto al resto de documentos de la colección.

$$A = \{(B:0{,}18),(C:0{,}15),(D:0{,}05),(E:0{,}04),(F:0{,}04),(G:0{,}05),(H:0{,}06)\}$$
$$B = \{(A:0{,}18),(C:0{,}31),(D:0{,}10),(E:0{,}05),(F:0{,}04),(G:0{,}03),(H:0{,}10)\}$$
$$C = \{(A:0{,}15),(B:0{,}31),(D:0{,}06),(E:0{,}04),(F:0{,}04),(G:0{,}03),(H:0{,}15)\}$$

Fig. 52 Vectores de similitudes para los puntos A, B y C de Fig. 51.

---

la revista *Communications of the ACM* entre 1.958 y 1.979. *CISI* incluye 1.460 artículos (título y resumen) compilados en el *Institute for Scientific Information*. *LISA* es una colección de 6.004 documentos (resúmenes) extraídos de la base de datos *Library and Information Science Abstracts* y, por último, *Cranfield* consta de 1400 documentos y es uno de los resultados del proyecto *Cranfield II* (véase página 140).

[1] `http://trec.nist.gov/`

[2] *Reuters Corpus* `<http://about.reuters.com/researchandstandards/corpus>` y *Reuters-21578* `<http://www.daviddlewis.com/resources/testcollections/reuters21578>`.

[3] Prescindiendo de los pares repetidos y aquellos que involucran al mismo documento.

A fin de determinar qué documentos exhiben un "comportamiento" análogo, esto es, tienen vectores de similitudes parecidos se ha empleado una medida que permite comparar las poblaciones de distintos ecosistemas, el denominado coeficiente de Bray-Curtis[1] (Bray y Curtis 1957, citado por Gauch 1982):

$$C_Z = \frac{2w}{a+b} \qquad (1)$$

En este coeficiente (véase la ecuación 1) *a* es la suma de las poblaciones de todas las especies en el primer ecosistema, *b* es la suma de las poblaciones en el segundo ecosistema y *w* es la suma de la población menor para cada especie presente en ambos ecosistemas (en Fig. 54 se muestra un ejemplo ilustrativo).

B={(*A*: 0,18), (*C*: 0,31), (*D*: 0,10), (*E*: 0,05), **(*F*: 0,04)**, **(*G*: 0,03)**, **(*H*: 0,10)**}
C={**(*A*: 0,15)**, (*B*: 0,31), **(*D*: 0,06)**, **(*E*: 0,04)**, **(*F*: 0,04)**, **(*G*: 0,03)**, (*H*: 0,15)}

a	0,18+0,31+0,10+0,05+0,04+0,03+0,10	**0,81**
b	0,15+0,31+0,06+0,04+0,04+0,03+0,15	**0,78**
w	0,15+0,06+0,04+0,04+0,03+0,10	**0,42**
C_Z	*2w/(a+b)*	**0,53**

Fig. 53 Cálculo del coeficiente de Bray-Curtis para los vectores de similitudes de los puntos B y C (véase Fig. 52).

Este coeficiente se puede aplicar al problema que nos ocupa de forma inmediata (véase Fig. 53) y permite obtener para cada pareja de documentos una nueva medida de similitud que es el producto de *PiRo* y C_Z. Posteriormente se determina un umbral adecuado para estos valores y aquellos pares de documentos cuya similitud $PiRo \cdot C_Z$ supere dicho umbral son incluidos en el mismo grupo.

Puesto que este algoritmo, véase Fig. 55, requiere $(n^2 - n)/2$ comparaciones tanto de documentos como de vectores de similitudes no es adecuado para clasificar colecciones demasiado grandes. Sin embargo, puesto que puede transformarse fácilmente en un algoritmo aglomerativo que emplea la matriz de similitudes $PiRo \cdot C_Z$ permite elaborar dendrogramas como los que se muestran en este capítulo y el anterior.

[1] Las principales características de este índice que lo hacen idóneo para este problema son dos: en primer lugar la ausencia del mismo componente en ambos vectores no se tiene en cuenta como "un punto en común" y en segundo los componentes mayores son los que dominan el coeficiente.

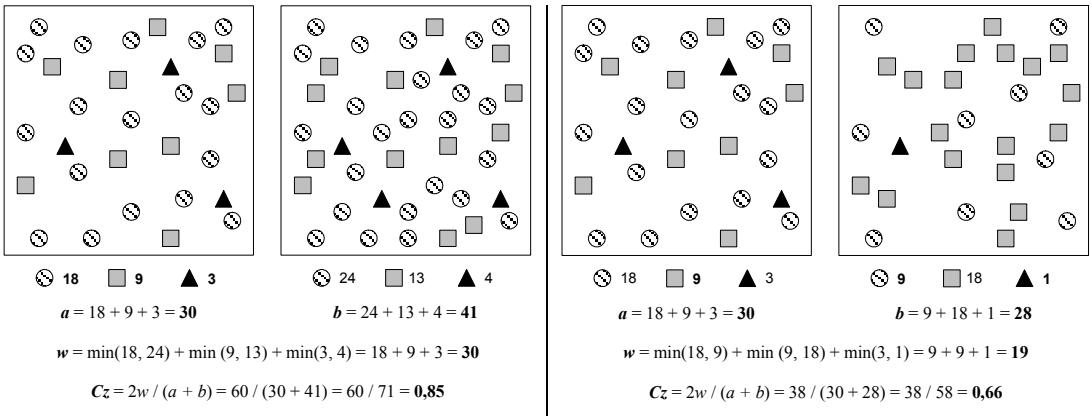


Fig. 54 Utilización del coeficiente de Bray-Curtis para la comparación de dos ecosistemas.

Se muestran aquí dos ejemplos de la evolución de un ecosistema en el que habitan 3 especies: los círculos rayados, los cuadrados y los triángulos. Los primeros son una especie vegetal que sirve de alimento a los cuadrados que son, a su vez, presa de los triángulos. A izquierda y derecha se muestra la evolución del ecosistema desde la misma situación de partida en la que hay 18 círculos, 9 cuadrados y 3 triángulos. En el caso de la izquierda la población de círculos ha pasado a contar con 24 individuos permitiendo el aumento de la población de cuadrados (13 individuos) y de triángulos (4 individuos). El coeficiente de Bray-Curtis señala que ambos ecosistemas son muy similares (0,85). En el caso de la derecha, en cambio, algo ha eliminado un número importante de triángulos lo que ha llevado a un aumento de la población de cuadrados (18) y una reducción en la población de círculos (9). El coeficiente de Bray-Curtis señala que se ha producido un cambio importante (0,66).

Algoritmo nonIncrementalBLClustering *(colección)*
Input: *colección*, una lista que contiene un vector de *n*-gramas por cada documento de la colección

1. **for each** (d_i, d_j) *en colección* **do**
2. $matriz(i)(j) \leftarrow (\mathbf{pi}(di, dj) + \mathbf{rho}(di, dj))/2$
3. $matriz(j)(i) \leftarrow matriz(i)(j)$
4. **loop**
5. **for each** (d_i, d_j) *en colección* **do**
6. $C_Z \leftarrow \mathbf{brayCurtis}(matriz(i), matriz(j))$
7. $parecidos(i)(j) \leftarrow matriz(i)(j) \cdot C_Z$
8. **loop**
9. $x \leftarrow \mathbf{media}(parecidos)$
10. $s \leftarrow \mathbf{desviacion}(parecidos)$
11. $umbral \leftarrow x + alfa \cdot s$
12. **for each** (d_i, d_j) *en colección* **do**
13. **if** $parecidos(i)(j) \geq umbral$
14. $clusters \leftarrow \mathbf{agrupar}(d_i, d_j)$
15. **end if**
16. **loop**
17. **return** *clusters*

Fig. 55 Algoritmo no incremental de clasificación automática.

2.2 Algoritmo incremental basado en *blindLight*

Cuando la colección de documentos es muy grande la utilización del algoritmo anterior puede ser enormemente costosa tanto temporal como espacialmente. Por ese motivo se ha desarrollado un algoritmo incremental que, a pesar de su sencillez, proporciona resultados adecuados al compararlo con otras técnicas de clasificación automática. En este caso, puesto que no se dispone de información completa acerca del "comportamiento" de cada documento en relación con el resto de elementos de la colección es inviable la aplicación del coeficiente de Bray-Curtis utilizado en el método anterior.

El procedimiento es relativamente sencillo (véase Fig. 56, Fig. 58 y Fig. 59). En primer lugar se obtiene el centroide de la colección a clasificar y la similitud de cada documento con el mismo. Una vez hecho esto se extraen aleatoriamente documentos individuales del conjunto original. Cada documento aleatorio es comparado con los grupos disponibles[1] obteniendo la similitud entre el documento y cada grupo. Hay que recordar que previamente se ha calculado la similitud de cada documento con el centroide de la colección y cada vez que se crea o modifica un grupo también se calcula la similitud de dicho grupo con el centroide.

Algoritmo incrementalBLclustering *(colección)*
Input: *colección*, una lista que contiene un vector de *n*-gramas por cada documento de la colección
1. *provisionales* ← ***tentativeClustering*** *(colección)*
2. *(singletons, provisionales)* ← ***separarSingletons*** *(provisionales)*
3. *(dispersos, definitivos)* ← ***testDispersion*** *(provisionales)*
4. *nuevos* ← ***tentativeClustering*** *(dispersos)*
5. *semilla* ← *definitivos + nuevos*
6. *clusters* ← ***tentativeClustering*** *(singletons, semilla)*
7. **return** *clusters*

Fig. 56 Algoritmo incremental de clasificación automática de documentos.

De este modo es posible determinar la similitud existente entre el documento y el grupo, y la media de las similitudes entre el centroide y el documento y entre el centroide y el grupo, respectivamente. Si la similitud entre el documento y el grupo es mayor que la segunda se asigna el documento al grupo (y se recalcula la similitud del mismo con el centroide) y en caso contrario se transforma el documento en un nuevo grupo (véase Fig. 57). Este proceso se repite hasta que la colección queda vacía.

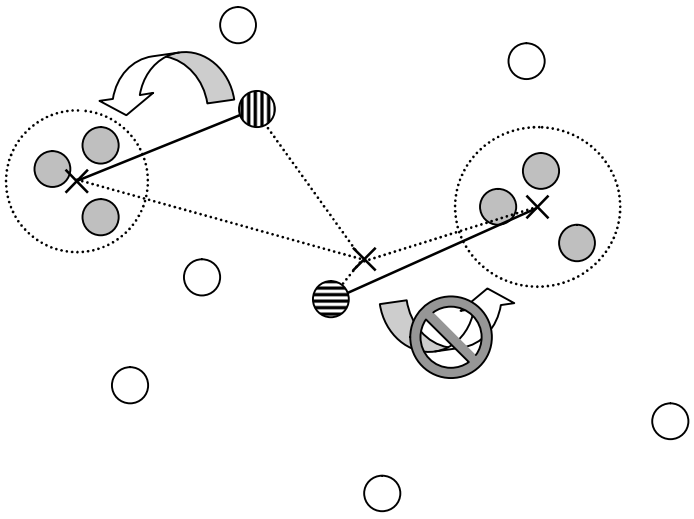


Fig. 57 Proceso de asignación de un documento a un grupo.

Un documento es asignado a un grupo si el parecido entre el grupo y el documento es superior a la media de las similitudes respectivas del grupo y el documento con el centroide de la colección. En la figura el documento rayado verticalmente puede asignarse al grupo de la izquierda mientras que el rayado horizontalmente no se asigna al de la derecha.

[1] Naturalmente, el primer documento extraído es convertido automáticamente en un grupo con un único documento.

Algoritmo tentativeClustering *(colección, semilla=λ)*
Input: *colección*, una lista que contiene un vector de *n*-gramas por cada documento de la colección

```
1.  if semilla ≠ λ
2.      clusters ← semilla
3.  end if
4.  centroide ← centroid (colección)
5.  for each documento d_i en colección do
6.      simDocs(i) ← ( pi (d_i, centroide) + rho (d_i, centroide))/2
7.  loop
8.  from i ← 0 to tamaño de colección do
9.      n ← random (tamaño de colección)
10.     if tamaño de clusters = 0 matriz(j)(i) ← matriz(i)(j)
11.         clusters(i)=d_n
12.         simClusters(i)=simDocs(n)
13.     else
14.         for each cluster k en clusters do
15.             sim ← ( pi (d_n, k) + rho (d_n, k))/2
16.             if sim > maxSim
17.                 maxSim ← sim
18.                 candidato ← k
19.             end if
20.         loop
21.         simDoc ← simDocs(n)
22.         simK ← simClusters(candidato)
23.         simAvg ← (simDoc + simK) / 2
24.         if maxSim > simAvg
25.             candidato ← agrupar (candidato, d_n)
26.             simClusters(candidato)←( pi (candidato, centroide) + rho (candidato,
    centroide))/2
27.         end if
28.     end if
29. loop
30. return clusters
```

Fig. 58 Algoritmo para obtener un conjunto de grupos "provisionales" (admite opcionalmente un conjunto de grupos "semilla").

Algoritmo testDispersion *(clusters)*
Input: *clusters*, un conjunto de grupos de documentos

```
1.  for each cluster k en clusters do
2.      variacionSims(k) ← coeficienteVariacion (k)
3.  loop
4.  x ← media (variacionSims)
5.  desv ← desviacion (variacionSims)
6.  umbral ← x + alfa ·desv
7.  for each cluster k en clusters do
8.      if variacionSims(k) > umbral
9.          coleccion ← extraerDocumentos (k)
10.     else
11.         clustersDef ← k
12.     end if
13. loop
14. return (coleccion, clustersDef)
```

Fig. 59 Algoritmo para determinar que grupos están "dispersos" y cuales son "definitivos".

En ese momento se dispone de una serie de grupos provisionales y, muy probablemente, de una serie de grupos constituidos por un único documento (*singletons*). Se aislan los primeros y se determina cuales constituyen grupos "dispersos". Para ello se calcula el coeficiente de variación de la similitud intra-grupal y, posteriormente, la media y desviación típica de dichos coeficientes para todos los grupos provisionales obteniendo un umbral. Aquellos grupos cuyo coeficiente de variación de similitud intra-grupal supere el umbral se consideran "dispersos" y sus documentos son trasladados a una nueva colección que es clasificada aplicando el procedimiento original. Los grupos obtenidos en esta clasificación son añadidos a los grupos provisionales no dispersos y todos pasan a considerarse grupos definitivos.

En este instante se dispone de un conjunto de grupos de documentos no dispersos y una serie de *singletons*. A fin de asignar, si es posible, dichos documentos aislados a un grupo se extrae el **medoide**[1] de cada grupo convirtiéndolo en un grupo por sí mismo y se aplica el algoritmo original sobre los *singletons* y estos grupos "semilla". Finalizada esta fase, algunos de los *singleton* habrán sido agrupados con algún medoide mientras que otros permanecerán aislados. Los que se han asociado a un medoide se integran en el grupo correspondiente mientras que los *singletons* se consideran grupos definitivos y el algoritmo finaliza con esta partición del conjunto original.

3 Algunos resultados de la aplicación de *blindLight* a la clasificación automática

A continuación se presentarán algunos resultados relevantes de la aplicación de *blindLight* a la clasificación automática de documentos. En primer lugar se describirá un experimento que permitió establecer una clasificación de distintos lenguajes naturales basándose tanto en datos léxicos como fonológicos. Seguidamente se mostrarán una serie de experimentos cuyos resultados dan soporte a la afirmación del autor acerca de la capacidad de *blindLight* para ofrecer resultados similares (y en algunos casos superiores) a los de métodos específicos.

3.1 Clasificación genética (y automática) de lenguajes naturales[2]

La mayor parte de lenguajes humanos están asignados a una familia que agrupa una serie de lenguas derivadas de un único idioma anterior. Un ejemplo que se cita con frecuencia son las lenguas romances que descienden del latín. Sin embargo, en muchos casos la lengua de origen es desconocida y es necesario reconstruir sus características a fin de determinar la evolución que siguió dicho idioma hasta producir lenguas conocidas. Este método, conocido como método comparativo, tiene sus orígenes en los estudios que realizó en el S. XIX August Schleicher sobre las lenguas indoeuropeas.

El método comparativo establece que dos idiomas están relacionados tan sólo si es posible reconstruir (aunque sea parcialmente) un idioma ancestro común. La razón es simple, tan sólo se consideran relaciones entre idiomas que han evolucionado por transmisión de padres a hijos durante generaciones. Así pues, la aplicación del método es lenta y compleja, y en muchos casos resulta imposible reconstruir un idioma anterior común, razón por la cual muchas lenguas humanas aparecen "aisladas" como únicos ejemplares de su propia familia (casos del vasco, el japonés o el coreano).

---

[1] El documento más parecido al centroide de un grupo.

[2] Gran parte de este apartado apareció en (Gayo Avello, Álvarez Gutiérrez y Gayo Avello 2004b).

Se han propuesto otras técnicas alternativas que también tienen como objetivo establecer vínculos de parentesco entre lenguas sin la necesidad de reconstruir ningún ancestro común y empleando tan sólo información léxica. Por ejemplo, la léxico-estadística (Swadesh 1950) o glotocronología (Lees 1953) y la comparación léxica masiva (Greenberg 1966).

La lexico-estadística trata de reconstruir árboles lingüísticos para lenguajes pertenecientes a la misma familia lingüística analizando el porcentaje de palabras afines[1] mientras que la glotocronología pretende, además, estimar la fecha en que dos lenguas divergieron a partir de un ancestro común. Por lo que se refiere a la comparación léxica masiva se basa en la comparación de palabras equivalentes (no necesariamente afines) en múltiples idiomas buscando sonidos similares en las mismas. Como principal éxito de esta última cabe destacar la clasificación por parte de Joseph H. Greenberg de los distintos idiomas africanos en cuatro grandes familias (Greenberg 1966); esta clasificación aunque inicialmente muy polémica es actualmente aceptada. Igualmente polémica es su clasificación de los idiomas nativos de América en tres grandes familias lingüísticas de las cuales la Amerindia no se admite generalmente como una familia única.

Es necesario decir que la mayor parte de los lingüistas no consideran estas técnicas en absoluto ortodoxas puesto que se basan únicamente en parecidos superficiales a nivel léxico y no fonético, por ejemplo, Poser y Campbell (1992) o Goddard y Campbell (1994). No obstante, las conclusiones obtenidas con dichas técnicas, en particular mediante la comparación léxica masiva, se han validado al aplicarse sobre idiomas cuya clasificación es bien conocida (caso de las lenguas indoeuropeas) o al proporcionar clasificaciones verificadas posteriormente de manera tradicional (caso de las lenguas africanas).

Así pues, y a pesar de la polémica, se han implementado distintos algoritmos que utilizan información meramente léxica para clasificar automáticamente distintos lenguajes humanos, por ejemplo (Dyen, Kruskal y Black 1992), (Kessler 1995), (Huffman 1998) o (Nerbonne y Heeringa 1997). Muchas de estas técnicas se basan en el clásico modelo vectorial, caso de Huffman que utiliza *Acquaintance* (Damashek 1995), o aplican la distancia de Levenshtein (1966) directamente sobre cadenas de caracteres (Kessler 1995) o fonemas (Nerbonne y Heeringa 1997). Recientemente Gray y Atkinson (2003) han analizado la evolución de las lenguas indoeuropeas con técnicas biocomputacionales y Warnow *et al.* (2004) y Evans, Ringe y Warnow (2004) han estudiado modelos estocásticos de evolución lingüística.

Por todo ello el autor de este trabajo consideró interesante la aplicación de *blindLight* al problema de la clasificación genética de lenguajes. La principal diferencia entre los trabajos anteriores y los experimentos que se llevaron a cabo radicó, además de en la técnica empleada, en el tipo de información lingüística utilizada. En primer lugar, se trabajó no sobre listas de palabras traducidas (como las listas de Swadesh[2]) sino sobre documentos

[1] También términos cognados. Palabras de distintos idiomas de significado y sonido similar. Como ejemplos de palabras afines pueden citarse *night*, *nuit* o *noche* en inglés, francés y español, respectivamente o *shalom* y *salaam* en hebreo y árabe.

[2] Una lista de Swadesh (1950) es una secuencia de palabras que constituyen un vocabulario básico (y presumiblemente antiguo). Todas las listas de Swadesh contienen el mismo vocabulario de tal modo que están alineadas a nivel de palabra constituyendo un *corpus* paralelo. Por ejemplo, la lista de Swadesh en castellano comienza con las palabras `{yo, tú/usted, él/ella, nosotros/nosotras, vosotros/vosotras, ellos/ellas, este, ese/aquel, …}` mientras la correspondiente al inglés contiene `{I, you/thou, he, we, you, they, this, that, …}`

completos[1]. Por otro lado además de información léxica también se utilizaron transcripciones fonéticas ya que ésta es una de las debilidades frecuentemente achacadas a las técnicas estadísticas.

Así, el primer experimento se realizó sobre vectores de *n*-gramas obtenidos a partir de los tres primeros capítulos del Libro del Génesis y, aplicando la versión no incremental y jerárquica del algoritmo de clasificación, se obtuvo un árbol con 14 idiomas (véase Fig. 60).

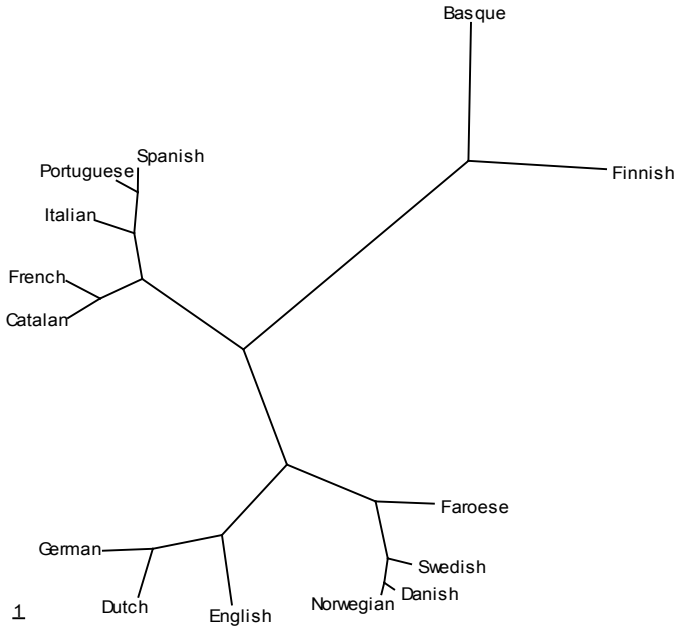


Fig. 60 Dendrograma mostrando las distancias entre muestras escritas de 14 idiomas europeos (tres primeros capítulos del Libro del Génesis).

En el segundo experimento se utilizaron vectores construidos a partir de transcripciones fonéticas (véase Fig. 61) de la fábula "El viento del norte y el sol" que se obtuvieron del *Handbook of the International Phonetic Association* (Manual de la Asociación Fonética International) (IPA 1999). Los idiomas que participaron en este último experimento fueron alemán, catalán, español, francés, gallego, holandés, inglés, portugués y sueco. En el árbol obtenido en esta ocasión se encontraban 8 idiomas presentes en el primer experimento (véase Fig. 62).

Al analizar los resultados obtenidos es necesario ser cauto, en primer lugar porque el autor no es lingüista y, en segundo, porque los lingüistas no dan demasiado crédito a los resultados obtenidos a partir de análisis estadísticos de datos puramente léxicos. No obstante, es preciso señalar que la segunda de las experiencias no ha empleado datos léxicos sino fonológicos producidos por expertos en el campo y los resultados obtenidos en dicha prueba son coherentes con los alcanzados en la primera.

Por otra parte, los resultados obtenidos en ambos casos clasifican de manera adecuada las lenguas indoeuropeas e incluso la estrecha relación manifestada (tanto léxica como fonéticamente) entre catalán y francés encuentra apoyo en ciertos autores, por ejemplo, Pere Verdaguer (1999).

---

[1] Recientemente el autor ha tenido noticia de un trabajo en el que utilizando la *Declaración Universal de los Derechos Humanos* se ha establecido una clasificación automática de 52 idiomas (Li *et al.* 2004, p. 3260-3261). La técnica empleada difiere de la utilizada por el autor pero los resultados son muy similares.

ðə ˈnɔɹθ ˌwɪnd ən ə ˈsʌn wɚ dɪsˈpjutɪŋ ˈwɪtʃ wəz ðə ˈstɹɑŋgɚ, wɛn ə ˈtɹævlɚ kem əˈlɑŋ ˈɹæpt ɪn ə ˈwɔɹm ˈklok. ðə əˈgɹid ðət ðə ˈwʌn hu ˈfɚst səkˈsidəd ɪn ˈmekɪŋ ðə ˈtɹævlɚ ˈtek ɪz ˈklok ˌɑf ʃʊd bi kənˈsɪdɚd ˈstɹɑŋgɚ ðən ðɪ ʌðɚ. ˈðɛn ðə ˈnɔɹθ ˌwɪnd ˈblu əz ˈhɑɹd əz hi ˈkʊd, bət ðə ˈmɔɹ hi ˈblu ðə ˈmɔɹ ˈklosli dɪd ðə ˈtɹævlɚ ˈfold hɪz ˈklok əˈɹaʊnd hɪm; ˌæn ət ˈlæst ðə ˈnɔɹθ ˌwɪnd ˌgev ˈʌp ði əˈtɛmpt. ˈðɛn ðə ˈsʌn ˈʃaɪnd ˌaʊt ˈwɔɹmli, ən ɪˈmidiətli ðə ˈtɹævlɚ ˌtʊk ˈɑf ɪz ˈklok. ən ˈso ðə ˈnɔɹθ ˌwɪnd wəz əˈblaɪʒ tɪ kənˈfɛs ðət ðə ˈsʌn wəz ðə ˈstɹɑŋgɚ əv ðə ˈtu.

The North Wind and the Sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew, the more closely did the traveller fold his cloak around him; and at last the North Wind gave up the attempt. Then the Sun shone out warmly, and immediately the traveller took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

Fig. 61 Transcripción fonética de la versión inglesa de la fábula de Esopo (*IPA* 1999, p. 44).

Debe quedar claro que en modo alguno se está aventurando ninguna hipótesis sobre la evolución de estos lenguajes, tan sólo se presentan unos resultados que, al ajustarse a clasificaciones construidas por humanos, y que junto con los obtenidos en el capítulo anterior para la clasificación de los mini-*corpora* paralelos apoyan la afirmación del autor acerca de que la utilización de *blindLight* como método de *clustering* permite obtener clasificaciones bastante "naturales" desde el punto de vista del usuario.

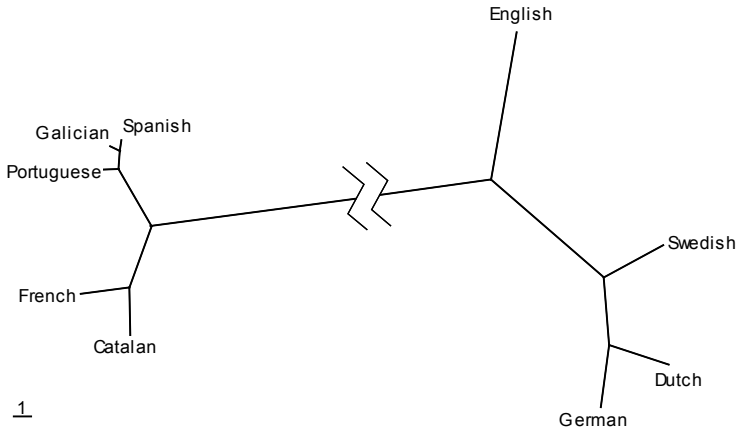


Fig. 62 Dendrograma mostrando las distancias entre muestras orales de 9 idiomas europeos (transcripciones fonéticas de la fábula "El viento del norte y el sol").

La distancia entre los sub-árboles Galo-Ibérico (izquierda) y Germánico es 23.985, más del doble de la distancia mostrada en la figura.

3.2 Comparación de *blindLight* con *SOM*

Ya se ha mencionado antes la aplicación de los Mapas Auto-Organizativos (*Self-Organizing Maps* o *SOM*) al problema de la clasificación automática de documentos. Este tipo de mapas, también denominados mapas de Kohonen (1982) por su inventor, pueden interpretarse como una distribución (generalmente bi o tridimensional) de neuronas situadas en posiciones fijas que se entrenan con vectores de características en un proceso competitivo. Para cada vector hay una única neurona ganadora que ajustará sus pesos para

aproximarse al vector de entrada. No obstante, el resto de neuronas también ajustan parcialmente sus pesos de forma inversamente proporcional a la distancia a que se encuentren de la vencedora. De este modo, se van vinculando los vectores a diferentes coordenadas del mapa y en caso de que estén etiquetados se asociarán sus etiquetas a las distintas zonas del mismo.

Ya el propio Kohonen propuso la utilización de mapas auto-organizativos para visualizar y explorar colecciones de artículos *USENET* (Honkela *et al.* 1996) –véase Fig. 63– y la misma técnica fue aplicada con objetivos similares a una serie de colecciones entre las que se incluía la colección *TIME* (Rauber y Merkl 1999) y la edición de 1990 de *The World Fact Book* de la *CIA* (Merkl y Rauber 1998).

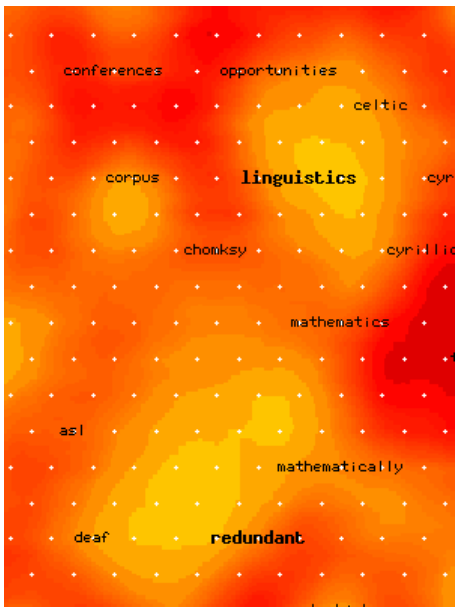


Fig. 63 Una zona de un mapa auto-organizativo para artículos publicados en los grupos `sci.lang.*`.

La colección *TIME* fue elaborada por Gerald Salton (1972) y consiste en 423[1] artículos publicados durante 1963 en la sección internacional de dicha revista. Los documentos abarcan toda una serie de temas tales como la situación en Vietnam, la guerra fría o escándalos políticos. El objetivo original de la colección fue el de servir de prueba estandarizada para la evaluación de sistemas de recuperación de información pero se ha utilizado con frecuencia para comparar distintos métodos de clasificación y categorización.

Por lo que respecta a *The World Fact Book* se trata de una publicación anual de la *CIA* que describe en términos geográficos, sociales, económicos y políticos distintas regiones y países del planeta. Merkl y Rauber (1998) utilizaron la edición de 1990 para elaborar una colección que consta de 246 documentos. Algunos ejemplos de documentos serían `argentin`, `soviet_u`, `gazastri`, `world` o `atlantic`.

En este apartado se describirán los resultados alcanzados al aplicar *blindLight* a la clasificación de ambas colecciones y se compararán con los obtenidos por Rauber y Merkl utilizando *SOM*[2] mientras que en el siguiente se comparará con técnicas como *k*-medias o *UPGMA*. Recuérdese que el autor afirma que esta técnica es capaz de ofrecer en todas sus aplicaciones, incluida la clasificación de documentos, resultados similares a los de técnicas específicas.

Al disponer de los grupos producidos por *SOM* la forma más sencilla de comparar esos resultados con los alcanzados con *blindLight* es calculando la similitud promedio. Para ello hay que calcular la media de las similitudes entre todos los posibles pares de

[1] Según Salton (1972, p. 5) la colección consta de 425 artículos pero las versiones actualmente disponibles sólo incluyen 423.

[2] *CIA World FactBook* <http://www.ifs.tuwien.ac.at/~andi/somlib/data/wfb90/wfb7a_1_1_0_0.html>, *TIME* <http://www.ifs.tuwien.ac.at/~andi/somlib/data/time60/time-map10x15_labels.html>

documentos de un grupo (Steinbach *et al.* 2000, p. 6) y posteriormente combinar estas medidas individuales en una medida única ponderando la similitud promedio de cada grupo en función del porcentaje de documentos de la colección incluidos en el mismo. Cuanto más elevada sea el valor de esta medida más cohesivos serán los grupos y, en consecuencia, mejor la técnica de clasificación.

En el caso de la primera colección se empleó la versión incremental de *blindLight* que dividió la colección *TIME* en 177 grupos de los cuales 78 estaban formados por un único documento. Para aplicar *SOM* Rauber y Merkl utilizaron un mapa bidimensional de 10x15 celdas[1]. De cara a la comparación de ambas técnicas se ha considerado cada celda como un grupo lo que supone 150 grupos de los que 41 fueron *singletons*. Los resultados obtenidos para esta colección aplicando *blindLight* y *SOM* se muestran en la Tabla 4.

blindLight		***SOM***	
Incluyendo *singletons*	Sin incluir *singletons*	Incluyendo *singletons*	Sin incluir *singletons*
0,597	**0,506**	0,547	0,498

Tabla 4. Similitud promedio de los resultados obtenidos con *blindLight* y *SOM* al clasificar la colección *TIME*.

Por lo que respecta a la segunda colección se empleó el método no incremental y se estableció sobre el dendrograma un punto de corte que proporcionase un número de grupos similar al obtenido con *SOM*. Al aplicar mapas auto-organizativos la colección queda dividida en 84 grupos de los cuales 26 estaban formados por un único documento. Los grupos obtenidos con *blindLight* fueron 86 incluyendo 24 *singletons*. Los resultados obtenidos en este segundo experimento se muestran en la Tabla 5.

blindLight		***SOM***	
Incluyendo *singletons*	Sin incluir *singletons*	Incluyendo *singletons*	Sin incluir *singletons*
0,731	**0,702**	0,712	0,678

Tabla 5. Similitud promedio de los resultados obtenidos con *blindLight* y *SOM* al clasificar la colección *CIA*.

Como se puede ver, los resultados obtenidos con *blindLight* son ligeramente mejores que los alcanzados por *SOM*. No obstante, el criterio general establece que una mejora de rendimiento es "apreciable" tan sólo si se encuentra entre el 5 y el 10% y "sustancial" si supera el 10% (Spärck-Jones 1974, citado por Rasmussen 2002) y no es ésta la situación. En el caso de la colección *TIME* hay una mejora de *blindLight* sobre *SOM* del 1,61% sin contar los *singletons* y del 9,14% contándolos. Por lo que respecta a la colección de la *CIA* el incremento en la similitud promedio es del 3,54% y del 2,67%, respectivamente. Así pues, se puede afirmar que las diferencias entre *blindLight* y *SOM* respecto a la clasificación de ambas colecciones no son relevantes.

Tales resultados son acordes con el objetivo de ofrecer una efectividad similar a los de técnicas específicas para una serie de tareas PLN por lo que permiten sostener dicha afirmación en lo que se refiere al problema de la clasificación automática y la técnica de mapas auto-organizativos. En el siguiente apartado se procederá a comparar *blindLight* con las técnicas de *k*-medias, *k*-medias bisecante (Steinbach, Karypis y Kumar 2000) y *UPGMA*[2] (Jain y Dubes 1988, citado por Zhao y Karypis 2002).

---

[1] El mapa está disponible en `http://www.ifs.tuwien.ac.at/~andi/somlib/data/time60/time-map10x15_labels.html`. Posteriormente los autores del mismo lo analizaron manualmente para reducir el número de grupos; no obstante, no se han empleado estos últimos debido a la intervención humana requerida para su construcción.

[2] *Unweighted Pair-Group Method with Arithmetic Mean.*

No obstante, no queremos concluir este apartado sin ofrecer algunas muestras de los resultados obtenidos con la nueva técnica propuesta a fin de permitir al lector valorar desde un punto de vista más práctico su utilidad. Después de todo, no se puede olvidar que el objetivo de la clasificación automática de documentos es señalar a un usuario final grupos con similitudes "interesantes".

Así, en Fig. 64 se presentan algunos grupos de documentos encontrados al clasificar con *blindLight* la colección *TIME* y que resultan particularmente representativos del tono de la colección. El primer grupo contiene una serie de documentos en torno a Nikita Khruschev, el segundo trata acerca de las elecciones en el Reino Unido, el siguiente sobre un escándalo político-sexual también en el Reino Unido y el último contiene documentos que tocan el problema de la tensión religiosa y política en Vietnam.

Cluster	T026, T032, T135, T248, T346, T539, **T542**, T558
Etiquetas	Khrushchev, Russia, Nikita, Soviet, party, Moscow, West, Kremlin, foreign…
Medoide	...Yet who should be serving up lemonade last week than that old realist **Nikita Khrushchev**. In the **Kremlin**'s marble-halled palace of the congresses, addressing the communist **party** central committee and more than 5,000 other comrades, **Nikita** promised that one great force would miraculously straighten out the **Soviet** economic mess: big chemistry... More important, he admitted that **Russia** would need credit and supplies, including entire factories, from the **West** but not, he fumed, at "fabulous profits" to the capitalists... Moreover, new products must show better design, because it is "no longer possible to tolerate" Russian consumer goods that "look less smart than **foreign** articles...

Cluster	**T105**, T240, T265, T270, T324, T430, T493, T497, T503, T512
Etiquetas	labor, party, minister, government, Britain, election, leader…
Medoide	...The **labor party** last week chose a new **leader** to carry its banner against the Tories in **Britain**'s coming general **election**... Some opposed his pro-common market views; others among **labor**'s intellectual center and right flinched at the thought of a working-class, up-from-the-ranks prime **minister**, and preferred to go to the country with an Oxford graduate and economics don like Wilson… The feuding has faded, and **labor** finds itself in the best shape in years to topple the **government** of Prime **Minister** Harold Macmillan...

Cluster	T170, T301, T315, **T337**, T342, T354
Etiquetas	Christine, Ward, Profumo, Keeler, British, government, girl, party, London, Britain, flat, Stephen, Russian…
Medoide	...The moral decay surrounding the **Profumo** affair, he tried hard to suggest, must be blamed on the Tories. Referring to **Christine Keeler**'s reported $14,000-a-week nightclub contract, Wilson declared: "there is something utterly nauseating about a system of society which pays a harlot 25 times as much as it pays its prime minister." For the rest, Harold Wilson stuck to the security issue and the **government**'s handling of the **Profumo** case, which he attacked as either dishonest or incompetent, or both... First, there was the **Christine**-**Profumo** affair itself, which, according to **Profumo**, lasted only a few months, from July to December 1961, but by other evidence possibly, lasted longer. During those same months, **Christine** also entertained **Russian** assistant naval attache Evgeny Ivanov, who had been pals for some time with her mentor, Dr. **Stephen Ward**...

Cluster	**T418**, T434, T464, T480, T498
Etiquetas	Diem, government, Viet, Saigon, Nhu, Buddhist, Nam, Dinh, south, Ngo, troops, army, Cong, brother, regime, war, president, Vietnamese, communist, crackdown, city, law, Mme, aid, martial, catholic, jail, roman, radio, Thuc …
Medoide	...to the **Diem government**, the **crackdown** obviously seemed necessary to protect the **regime** and enforce the **law** of the land against **Buddisht** defiance... it also put U.S. policy in **south Viet Nam**, which involves the lives and safety of 14,000 U.S. **troops**, into an agonizing dilemma... **roman catholic Diem** may finally have shattered his own political usefulness. he also opened up the possibilities of coups, counter-coups, and even civil **war** from all of which only the **communist Viet Cong** could benefit... Pagodas, sporting protest signs in **Vietnamese** and English... appeals for **aid** were broadcast to **President** Kennedy... At a grisly, well-organized press conference in **Saigon**... his **brother** and sister-in-**law**, **Ngo Dinh Nhu** and **Mme**... ten truckloads of bridge defenders were carted away to **jail**... and an estimated 500 people were arrested throughout the **city**... under the **martial law** proclamation, the **army** was given blanket search-and-arrest powers and empowered to forbid all public gatherings...

Fig. 64 Algunos grupos obtenidos al clasificar la colección *TIME*.

Para cada grupo se muestran los documentos que forman parte del mismo, el medoide del grupo (en negrita y extractado) y una serie de palabras que etiquetan el grupo obtenidas automáticamente al extraer aquellas que aparecen en un porcentaje elevado de documentos del grupo.

En Fig. 65 se muestran sobre una serie de mapas algunos de los grupos encontrados al analizar la colección de documentos de la *CIA* con *blindLight* (como se recordará, dicha colección está básicamente constituida por descripciones de países). Nótese cómo los grupos han incluido parámetros tanto geográficos (países africanos y americanos) como, en muchos casos, políticos (grupos localizados en oriente medio, el grupo EEUU-URSS), socio-económicos (el grupo formado por RFA, Suiza y Austria) y/o ideológicos (países del "Telón de Acero").

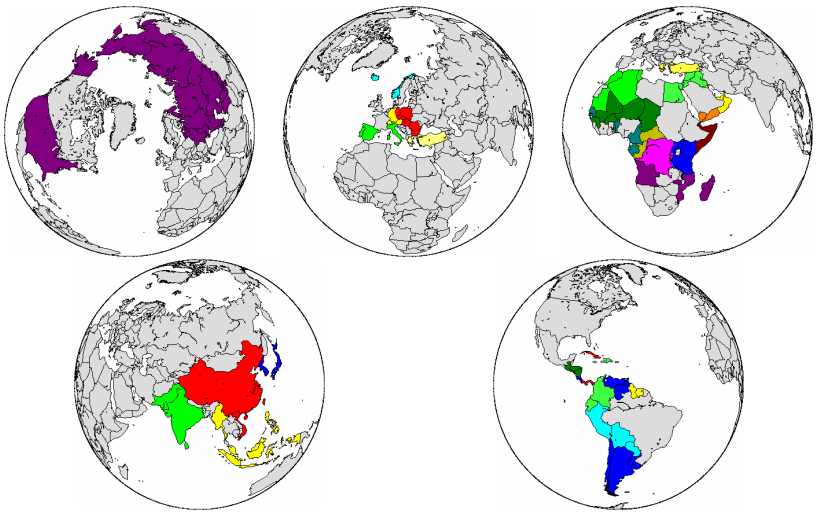


Fig. 65 Grupos de regiones localizados por *blindLight* analizando los textos de la *CIA*.

(De izquierda a derecha y de arriba abajo. Recuérdese que los datos son previos al desmoronamiento del bloque comunista). El primer grupo incluye tan sólo a EEUU y la URSS. El segundo mapa muestra a Portugal, España e Italia en un grupo, los países "germánicos" (RFA, Suiza y Austria) en otro, Grecia, Chipre y Turquía en un tercero, el "Telón de Acero" como cuarto grupo (señalado en rojo) y dos países nórdicos en un quinto. En tercer lugar se muestra una segmentación de África que parece responder a criterios geográficos, resulta interesante el grupo constituido por Egipto, Israel, Siria, Jordania, Líbano e Iraq. El siguiente gráfico divide Asia en cuatro grupos. El último mapa muestra una segmentación de América central y del sur por criterios fundamentalmente geográficos, nótese la ausencia de Brasil.

En la Tabla 6 se presentan algunos grupos particularmente interesantes y un análisis *post mortem* del "criterio" subyacente a cada clasificación. En Fig. 66 se muestran otros grupos obtenidos al "cortar" el dendrograma en un nivel superior, aquí tienen mayor peso los factores geográficos (fundamentalmente en Asia, África y América) y económicos (el primer mapa muestra los países más industrializados).

Por lo que respecta a grupos que contienen un único documento se han encontrado, en el "primer corte", Vaticano, Libia, Mongolia, Nigeria, Etiopía, Sudán, Ghana, Malta, Haití, México, Brasil, Arabia Saudí, Tailandia, Francia, S. Kitts, Georgia del Sur y las Islas Sandwich, Gibraltar, los territorios franceses en el Antártico, la Antártida y Jan Mayen (un territorio noruego). En cuanto al "segundo corte" los *singletons* fueron el Vaticano, Libia y Nigeria. Dada la naturaleza de la colección estos grupos podrían interpretarse como regiones "aisladas" de algún modo de sus vecinos, lo cual parece especialmente aplicable en los casos del Vaticano (único estado monárquico-sacerdotal del mundo), la Antártida o Gibraltar (único territorio colonial en suelo europeo).

Grupo	World, Svalbard, Gaza Strip, West Bank
Post mortem	El primer documento describe la situación mundial en términos generales. Svalbard es un territorio de soberanía noruega pero explotable por otros países, con presencia soviética y disputas marítimas entre ambos países. La franja de Gaza y Cisjordania son territorios ocupados por Israel y sin una situación política definida.

Grupo	Arctic ocean, Atlantic ocean, Indian ocean, Pacific ocean
Post mortem	Océanos

Grupo	Andorra, San Marino, Liechtenstein, Monaco
Post mortem	Microestados con relaciones especiales con otras naciones, en particular en términos de defensa: Liechtenstein con Suiza, Mónaco con Francia y Andorra con España y Francia.

Grupo	Iraq-Saudi Aurabia Neutral Zone, Paracel Islands, Spratly Islands
Post mortem	Territorios sin presencia humana permanente cuya defensa es responsabilidad de varios países o cuya soberanía es disputada por varios países.

Grupo	American Samoa, Guam, Northern Mariana Islands, Puerto Rico, Virgin Islands
Post mortem	Islas que constituyen territorios de EEUU (American Samoa, Guam y Virgin Islands) o estados asociados a EEUU (Northern Mariana Islands y Puerto Rico) y cuya defensa es responsabilidad de este pais.

Grupo	Marshall Islands, Federated States of Micronesia, Palau
Post mortem	Estados libres asociados a EEUU, situados en el Pacífico y cuya defensa es responsabilidad de este pais.

Grupo	Navassa Island, Kingman Reef, Palmyra Atoll
Post mortem	Territorios deshabitados de EEUU.

Grupo	Falkland Islands (Malvinas), St. Helena
Post mortem	Islas que constituyen territorios dependientes del Reino Unido y cuya defensa es responsabilidad de este país.

Grupo	Ashmore and Cartier Islands, Coral Sea Islands
Post mortem	Islas que constituyen territorio de Australia y carecen de población autóctona aunque pueden estar pobladas estacionalmente.

Grupo	Clipperton Island, Bassas da India, Europa Island, Tromelin Island, Glorioso Islands, Juan de Nova Island
Post mortem	Islas deshabitadas posesión de Francia.

Grupo	French Guiana, Reunion, Guadeloupe, Martinique, Mayotte, St. Pierre and Miquelon
Post mortem	Territorios y “colectividades” francesas de ultramar.

Grupo	French Polynesia, New Caledonia, Wallis and Futuna
Post mortem	Territorios franceses de ultramar situados en el Pacífico.

Tabla 6. Grupos producidos por *blindLight* y análisis *post mortem* de los mismos.

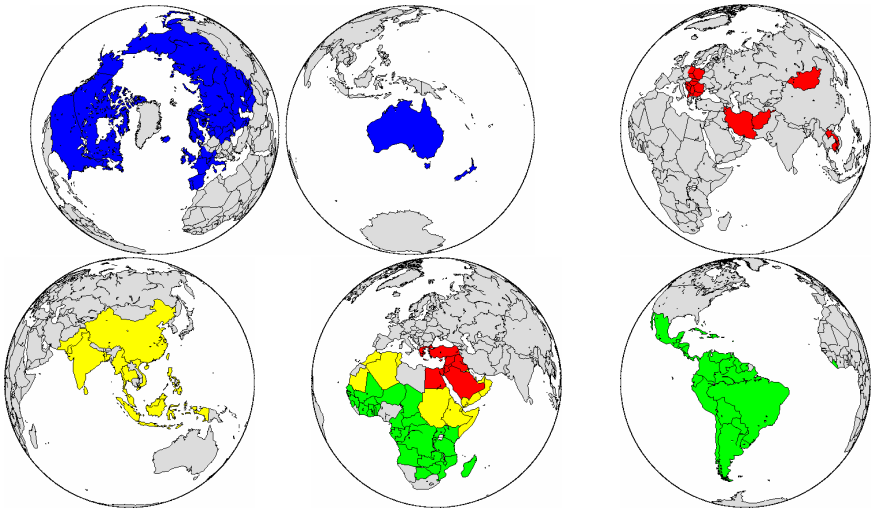


Fig. 66 Grupos de regiones obtenidos al "cortar" el dendrograma en un nivel superior.

El primer gráfico muestra el denominado "Norte Rico" que incluye a la URSS, Australia y Nueva Zelanda. El segundo mapa contiene países vinculados a la URSS ideológica, económica, militar y/o geográficamente: el "Telón de Acero", Irán (fronterizo) o Afganistan (fronterizo y ocupado). El tercer gráfico agrupa la mayor parte de Asia mientras el siguiente divide África en tres grandes grupos, destacando nuevamente la zona de Oriente Próximo. El último mapa coincide con Latinoamérica (nótese la ausencia de la Guyana Francesa) aunque incluye Liberia en un curioso salto transatlántico.

3.3 Comparación de *blindLight* con *k*-medias, *k*-medias bisecante y *UPGMA*

En el apartado anterior se han presentado los resultados obtenidos al aplicar *blindLight* y *SOM* sobre dos colecciones de documentos. Dichos resultados mostraron que las diferencias entre ambas técnicas no son relevantes. En este apartado se aprovechará el trabajo llevado a cabo por Steinbach, Karypis y Kumar (2000) para comparar la técnica propuesta por el autor con otras más "tradicionales".

En su trabajo Steinbach *et al.* presentaron un estudio experimental acerca del rendimiento de distintas técnicas de clasificación automática de documentos, en particular *k*-medias, una modificación de la misma denominada *k*-medias bisecante y un método jerárquico y aglomerativo clásico, *UPGMA*. El objetivo fundamental de dicho trabajo era determinar si, efectivamente, los métodos jerárquicos producen mejores clasificaciones que los métodos particionales. Para ello, utilizaron una serie de colecciones de documentos y obtuvieron un conjunto de medidas de la "calidad" de los resultados a fin de comparar las distintas técnicas. Del mismo modo, aplicando *blindLight* sobre una o más de dichas colecciones podrían obtenerse resultados análogos y comparar la técnica propuesta por el autor con las anteriores.

No obstante, se dan toda una serie de circunstancias que hicieron muy difícil la utilización de la mayor parte de las colecciones empleadas por Steinbach *et al.* En primer lugar, aun cuando los datos están accesibles en el sitio web de uno de los autores[1] no se ofrecen como texto plano sino procesados para su utilización con el paquete de *software CLUTO* haciéndolos inútiles para *blindLight* (véase Fig. 67).

[1] `http://www-users.cs.umn.edu/~karypis/cluto/files/datasets.tar.gz`

```
                                                    edition
                                                    dewei
4663 41681 83181                                    decim
1430 1 476 1 514 1 38 1 1024 1                      classif
13255 1 8549 1 2460 1 4987 1 175 1 249 1            studi
2186 1 1279 1 182 1 257 1 4515 1                    histori
…                                                   decimalclassif
                                                    ddc
                                                    …
```

Fig. 67 Datos procesados para ser empleados por *CLUTO*.

Las colecciones de documentos se representan mediante vectores de términos (a la izquierda). Cada fila del archivo se corresponde al vector de un documento donde los términos han sido reemplazados por índices enteros. A la derecha se muestran algunos de los términos empleados en la colección, su posición dentro del fichero sirve como índice del término para la representación vectorial. Este tipo de representación resulta inútil para *blindLight* que trabaja sobre el texto plano original de los documentos.

Por otro lado, al intentar localizar los textos originales se comprobó que varias de las colecciones son particiones utilizadas en la *Text REtrieval Conference* (*TREC*) que ni son libres ni gratuitas. Aún peor, la descripción para obtener dos particiones de la colección *Reuters-21578* es incompleta y el autor fue incapaz de reproducirlas aun disponiendo de ella. La única colección que pudo encontrarse y comprobarse que se correspondía con la descripción dada por Steinbach *et al.* fue la denominada *wap*[1] (*WebACE Project*).

Así pues, se aplicó *blindLight* (en su versión incremental) sobre esta colección (Han *et al.* 1998) que consta de 1560 páginas web extraídas de *Yahoo!* y asignadas a una única categoría de 20 posibles. Esto permitió no sólo que se pudiese calcular la similitud promedio (véase pág. 84) de los resultados obtenidos sino también la entropía de los mismos. Sin embargo, es necesario decir que en la bibliografía examinada se han encontrado dos definiciones de entropía distintas, Zhao y Karypis (2002) y Steinbach *et al.* (2000), y que a lo largo de este apartado se utilizará la segunda definición a fin de poder comparar adecuadamente los resultados obtenidos por Steinbach *et al.* y los alcanzados con *blindLight*.

Según Zhao y Karypis (2002) dado un grupo S_r de tamaño n_r la entropía del mismo se define como:

$$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}$$

donde q es el número de clases en la colección, n^i_r es el número de documentos de la clase i-ésima que fueron asignados al grupo r-ésimo. La entropía de la clasificación final se define como la suma de las entropías de todos los grupos ponderadas de acuerdo a su tamaño, es decir:

$$Entropia = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r)$$

En cambio, según Steinbach *et al.* (2000, p. 7) la entropía de un grupo individual sería:

$$E(S_r) = -\sum_{i=1}^{q} p_{ir}\log p_{ir} = -\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}$$

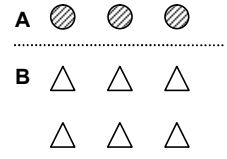
[1] `ftp://ftp.cs.umn.edu/dept/users/boley/PDDPdata/doc-K/`

Como se puede ver, la diferencia entre ambas definiciones es mínima y tan sólo cambia los valores numéricos y no la interpretación de la medida: a menor entropía mejor clasificación (véase Fig. 68).

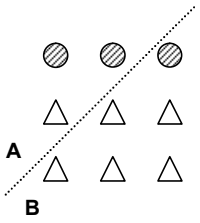
En las siguientes tablas se muestran los resultados obtenidos por *blindLight* para la colección *wap* y se comparan con los obtenidos por otras técnicas de clasificación (Steinbach *et al.* 2000, p. 14 y 15). Es necesario señalar que mientras dichas técnicas requieren la especificación del número de grupos a encontrar, *k*, la técnica propuesta por el autor no requiere tal parámetro por lo que los datos ofrecidos para *k*-medias, *k*-medias bisecante y *UPGMA* se han proyectado para *k*=70 (el número de grupos en que *blindLight* parte la colección) a partir de los publicados originalmente para *k*=16, 32 y 64.

blindLight	*k-medias*	*k-medias bisecante*	*k-medias bisecante "refinado"*	*UPGMA*	*UPGMA "refinado"*
1,1907	1,2230	**1,0888**	**1,0397**	1,3486	1,2561
Diferencia respecto a *blindLight*	-2,64% Inapreciable (A favor de *bL*)	9,36% Apreciable	14,52% Sustancial	-11,71% Sustancial (A favor de *bL*)	-5,21% Apreciable (A favor de *bL*)

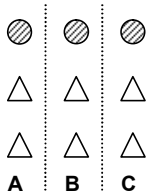
Tabla 7. Entropía de las distintas clasificaciones de la colección *wap*.



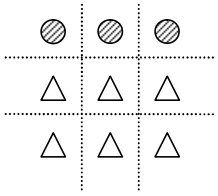
$$E(A) = -\left(\tfrac{3}{3}\cdot\log\tfrac{3}{3} + \tfrac{0}{3}\cdot\log\tfrac{0}{3}\right) = -(1\cdot\log 1 + 0\cdot\log 0) = 0$$
$$E(B) = -\left(\tfrac{0}{6}\cdot\log\tfrac{0}{6} + \tfrac{6}{6}\cdot\log\tfrac{6}{6}\right) = -(0\cdot\log 0 + 1\cdot\log 1) = 0$$
$$E = 0$$



$$E(A) = -\left(\tfrac{2}{3}\cdot\log\tfrac{2}{3} + \tfrac{1}{3}\cdot\log\tfrac{1}{3}\right) = -(-0{,}117 - 0{,}159) = 0{,}276$$
$$E(B) = -\left(\tfrac{1}{6}\cdot\log\tfrac{1}{6} + \tfrac{5}{6}\cdot\log\tfrac{5}{6}\right) = -(-0{,}130 - 0{,}066) = 0{,}196$$
$$E = \tfrac{3}{9}\cdot 0{,}276 + \tfrac{6}{9}\cdot 0{,}196 = 0{,}223$$



$$E(A) = E(B) = E(C) =$$
$$= -\left(\tfrac{1}{3}\cdot\log\tfrac{1}{3} + \tfrac{2}{3}\cdot\log\tfrac{2}{3}\right) = -(-0{,}159 - 0{,}117) = 0{,}276$$
$$E = 0{,}276$$



$$E(A) = E(B) = \ldots = E(H) = E(I) =$$
$$= -\left(\tfrac{1}{1}\cdot\log\tfrac{1}{1} + \tfrac{0}{1}\cdot\log\tfrac{0}{1}\right) = -(1\cdot\log 0 + 0\cdot\log 0) = 0$$
$$E = 0$$

Fig. 68 Cálculo del valor de la entropía para cuatro soluciones de agrupamiento en relación con una clasificación externa.

Cuanto mayor es la semejanza de la solución obtenida y la clasificación externa menor es la entropía de dicha solución. El valor mínimo posible es 0 que supondría una clasificación idéntica a la externa o bien una solución trivial consistente en la división del conjunto de elementos en grupos formados por un único *ítem* (último caso).

blindLight	*k-medias*	*k-medias bisecante*	*k-medias bisecante "refinado"*	*UPGMA*	*UPGMA "refinado"*
0,4270	0,3943	0,3914	0,3988	0,3634	0,3728
Diferencia respecto a *blindLight*	8,29% Apreciable	9,10% Apreciable	7,07% Apreciable	17,50% Sustancial	14,54% Sustancial

Tabla 8. Similitud promedio de las distintas clasificaciones de la colección *wap*.

Como se puede ver en las tablas anteriores, hay dos técnicas (*k*-medias bisecante y *k*-medias bisecante refinado) que obtienen una entropía menor que *blindLight* y tres técnicas que obtienen peores resultados que la técnicas propuesta por el autor. No obstante, tan sólo hay una técnica que mejora sustancialmente los resultados obtenidos por *blindLight*. Por lo que se refiere a la similitud promedio (o lo que es lo mismo, la cohesión) de los grupos en que *blindLight* divide la colección parece ser ligeramente mejor que la de las soluciones encontradas por otros algoritmos.

En resumen, a la vista de los resultados obtenidos en los experimentos descritos en este apartado y en el anterior puede concluirse que, al menos en lo que se refiere a las colecciones *TIME*, *CIA* y *wap*, al aplicar la técnica propuesta por el autor al problema de clasificar automáticamente una colección de documentos es posible obtener unos resultados semejantes, si no mejores, que los de técnicas específicas como mapas auto-organizativos, métodos particionales y métodos jerárquicos.

4 Influencia del tamaño de los *n*-gramas en la clasificación

Al describir los experimentos anteriores no se ha hecho mención alguna al tamaño de *n*-grama utilizado. Así, en el caso de las colecciones *TIME* y *CIA* se utilizaron 4-gramas mientras que para la colección *wap* se emplearon 2-gramas. Las razones que llevaron a utilizar tamaños diferentes fueron de índole práctica: un texto cualquiera de 700 palabras (alrededor de 3400 caracteres) contiene más de 400 2-gramas diferentes, alrededor de 1.400 3-gramas y más de 2.000 4-gramas. Obtener tales vectores con sus correspondientes significatividades así como combinarlos para calcular los valores de Π y P puede resultar muy costoso para colecciones grandes. Por tanto, es necesario determinar qué influencia tiene el tamaño de *n*-grama en los resultados obtenidos por *blindLight* al clasificar un conjunto de documentos a fin de evaluar en cada caso cuál es el más idóneo.

A este fin se prepararon dos subconjuntos de las colecciones *CIA* y *wap* que contenían 50 y 156 documentos seleccionados al azar lo que suponía, respectivamente, el 20% y el 10% de las colecciones originales. En ambos casos se obtuvieron vectores de 2-, 3- y 4-gramas para los documentos y se aplicó el algoritmo de clasificación no incremental[1] a cada uno de los seis conjuntos de datos obtenidos.

En el caso del subconjunto de la colección *CIA*, a partir de ahora *CIA-50*, se procedió a calcular el valor de la similitud promedio para distintas clasificaciones obtenidas con cada una de las tres versiones obteniéndose el gráfico que se muestra en Fig. 69.

Este gráfico parece sugerir que al utilizar 3-gramas para construir los vectores de documentos se obtienen clasificaciones con una similitud promedio superior, por tanto más cohesivas y, en teoría, preferibles. No obstante, puesto que para cada versión de la colección

[1] Debido al pequeño tamaño de estos subconjuntos es posible utilizar el algoritmo no incremental que siempre produce la misma clasificación. No obstante, como se verá después, las agrupaciones obtenidas con el algoritmo incremental y con el algoritmo no incremental son básicamente equivalentes.

CIA-50 (esto es, para cada tamaño de *n*-grama) se había obtenido una clasificación diferente era posible evaluar su cohesión sobre cada una de las versiones de los datos. Por ejemplo, dada la clasificación obtenida a partir de los vectores de 2-gramas es posible calcular la similitud promedio no sólo para dichos vectores sino también para los vectores de 3- y 4-gramas aun cuando éstos no se hubiesen utilizado para obtener dicha clasificación. Lo mismo es aplicable a las otras dos clasificaciones.

Al hacer esto se comprobó (véase Fig. 70) que las tres clasificaciones obtenían aproximadamente los mismos valores de similitud promedio cuando eran evaluadas sobre la misma colección de vectores para la colección (esto es, para el mismo tamaño de *n*-grama). No obstante, deducir de esto que las clasificaciones obtenidas son equivalentes parece arriesgado al tratarse la similitud promedio de una medida de calidad "interna" (Steinbach *et al.* 2000, p. 6).

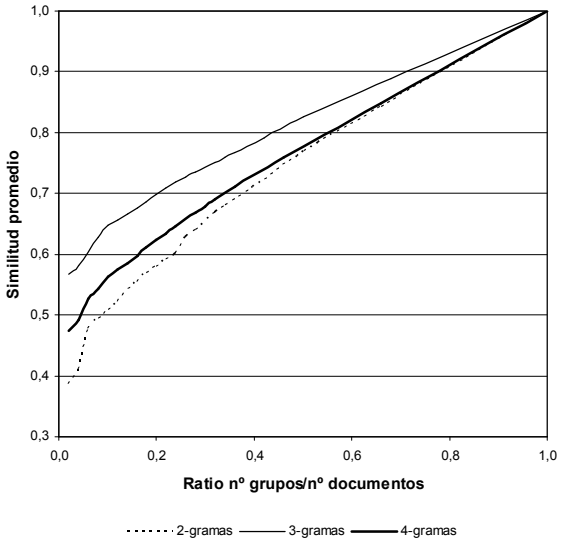


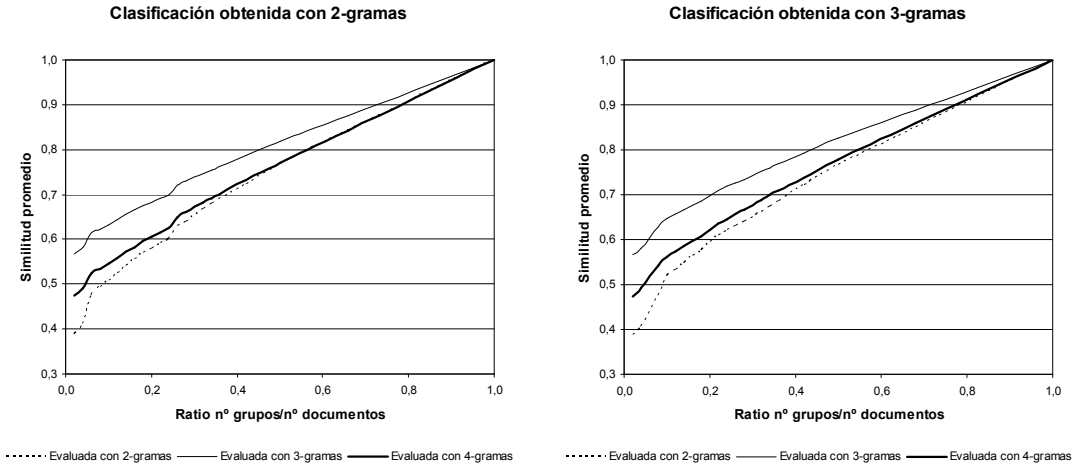
Fig. 69 Similitud promedio obtenida para cada una de las versiones de la colección *CIA-50* y clasificaciones con distinto números de grupos.

Por este motivo resultó tremendamente interesante el subconjunto *wap-156* puesto que se disponía de una clasificación externa respecto a la cual calcular la entropía (Steinbach *et al.* 2000, p. 7) y la pureza[1] (Zhao y Karypis 2002, p. 11) de las clasificaciones obtenidas aplicando *blindLight*. Los resultados obtenidos se muestran en Fig. 71. Como se puede ver, a medida que se emplean *n*-gramas de mayor tamaño para construir los vectores las clasificaciones obtenidas tienen menor entropía y una pureza superior. Las diferencias respecto al uso de 2-gramas son sensibles con independencia del número de grupos obtenidos y en el caso de 3-gramas y 4-gramas disminuyen cuando el número de grupos es del orden del número de documentos.

Así pues, como era de esperar a medida que aumenta el tamaño de *n*-grama utilizado mejora la calidad de la clasificación obtenida. Además de esto, se realizó un experimento con el subconjunto de la colección *wap* y el algoritmo incremental a fin de determinar si existen diferencias sensibles en los resultados obtenidos con ambas versiones del método. Puesto que el método incremental es estocástico se realizaron 20 ejecuciones del mismo sobre los vectores de 4-gramas (véase Tabla 9). Estos datos fueron promediados y representados junto con los resultados del algoritmo no incremental (véase Fig. 72) llegándose a la conclusión de que ambas versiones son básicamente equivalentes.

En conclusión, la técnica *blindLight* puede emplearse como método de clasificación automática de documentos obteniéndose resultados comparables e incluso mejores que los obtenidos al aplicar técnicas específicas.

[1] Recuérdese que la pureza evalúa en qué medida un grupo de una clasificación contiene documentos de una única clase.



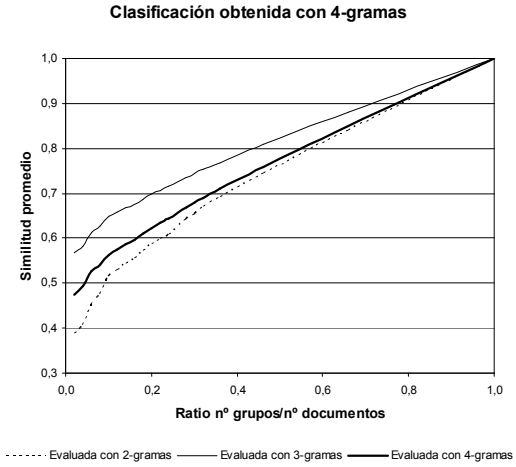


Fig. 70 Similitud promedio de las tres clasificaciones evaluadas sobre las tres versiones de la colección.

Grupos obtenidos	Entropía	Pureza
15	1,6040	0,4423
16	1,4882	0,4936
16	1,5042	0,4808
16	1,4104	0,4936
17	1,0654	0,5962
17	1,5967	0,4167
17	1,4087	0,4872
17	1,4694	0,5000
17	1,5058	0,4808
17	1,4620	0,4808
18	1,3639	0,4936
18	1,4449	0,4744
18	1,2991	0,5256
19	1,2248	0,5513
19	1,3884	0,4744
19	1,1151	0,6029
19	1,3815	0,5000
19	1,5627	0,4359
20	1,3761	0,4936
20	1,0529	0,5897

Tabla 9. Resultados obtenidos al ejecutar 20 veces el algoritmo incremental sobre el subconjunto de la colección *wap-156* elaborada con 4-gramas.

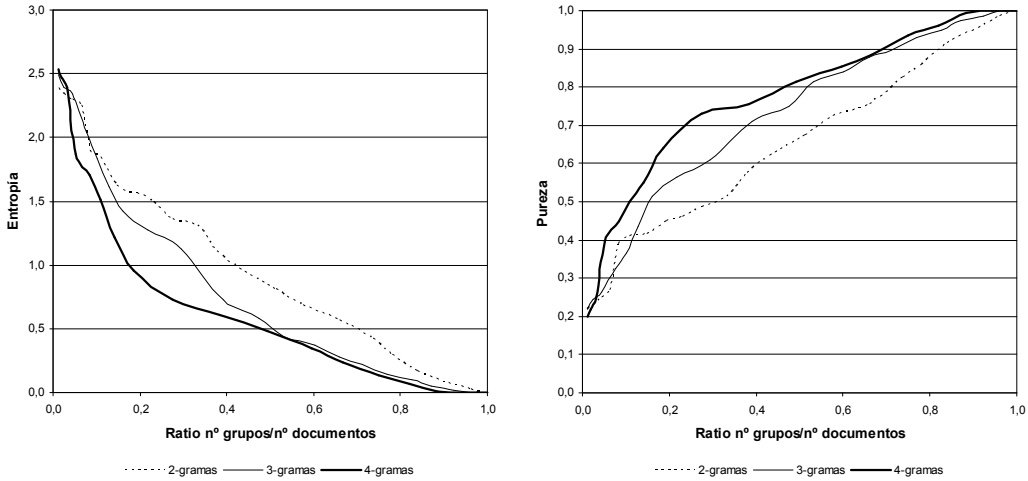


Fig. 71 Entropía y pureza de las clasificaciones obtenidas al utilizar distintos tamaños de *n*-grama en la clasificación de la colección *wap-156*.

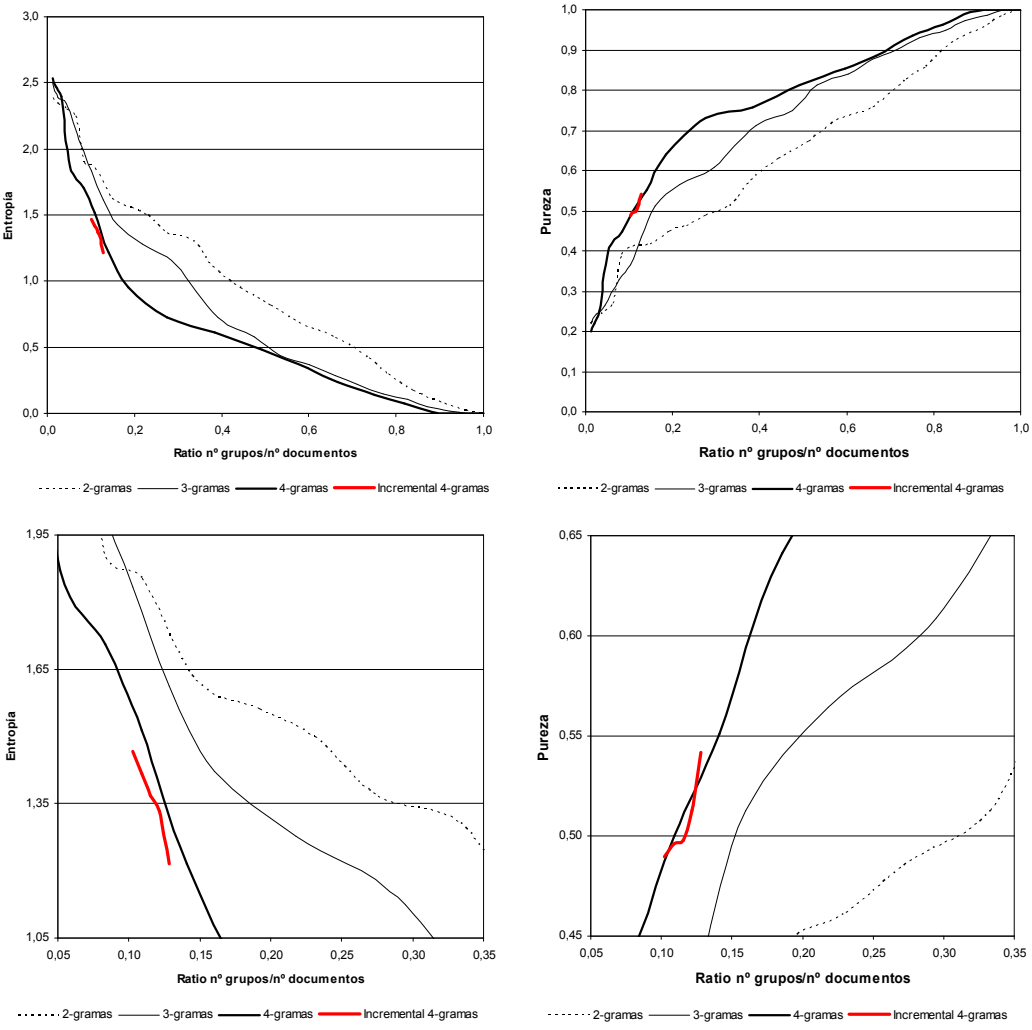


Fig. 72 Resultados de entropía y pureza obtenidos por el algoritmo incremental comparados con los obtenidos por la versión no incremental para la colección *wap-156*.

CAPÍTULO 5

CATEGORIZACIÓN DE DOCUMENTOS CON *BLINDLIGHT*

La categorización de documentos es el proceso por el cual se asocian una o más categorías a textos escritos en un lenguaje natural basándose tan sólo en su contenido. Aunque es posible construir de manera "manual" un categorizador, las técnicas estadísticas y, por tanto, automáticas son actualmente las preferidas puesto que no sólo ofrecen un rendimiento muy adecuado sino que resulta mucho más sencillo seleccionar un conjunto de ejemplos para entrenar un algoritmo que elaborar reglas manualmente. Así pues, esta tarea entra dentro del campo del aprendizaje automático y es posible aplicarle una gran variedad de técnicas disponibles. En este capítulo se presentará la categorización de documentos en tanto que problema de aprendizaje, se mostrarán algunas de las posibles aplicaciones de la categorización automática de texto libre, se revisarán las distintas técnicas aplicadas al problema y, por último, se describirá la forma de utilizar blindLight como categorizador y se presentarán los resultados obtenidos con esta nueva técnica.

1 Categorización automática de documentos

Según el diccionario de la *RAE* (2001) la categorización es la acción y efecto de organizar o clasificar mediante categorías, entendidas éstas como un elemento de clasificación. La definición no es muy clara pero no parece prudente adentrarse en terreno filosófico y sí en cambio proponer una definición de categorización útil en el campo del tratamiento de información. Así, podría definirse la categorización como la acción que realiza un agente al etiquetar *ítems* con una o más categorías de un conjunto predefinido basándose en las características de los mismos.

La categorización así entendida se viene utilizando desde hace siglos. Aristóteles[1] o Linneo, por ejemplo, establecieron categorías de plantas y animales. Posteriormente, se

[1] Aristóteles fue quien acuño el término "categoría" a partir del término *katigoro* (κατηγορώ), "acusar", "predicar" de algo o alguien.

desarrollaron diversos métodos para organizar los recursos almacenados en bibliotecas como el de la Biblioteca del Congreso de EE.UU. (*Library of Congress Classification*, *LCC*) o el Sistema Decimal Universal (*Universal Decimal Classification*, *UDC*). Más reciente es el sistema de categorización de la *ACM* (*ACM Computing Classification System*) para clasificar artículos sobre informática y otros para categorizar recursos web como por ejemplo la jerarquías desarrolladas por *Yahoo!* (Steinberg 1996) o al amparo del *Open Directory Project* (*ODP*).

La mayor parte de estos sistemas estructuran las categorías en jerarquías lo cual es en ocasiones un inconveniente[1] (véase Fig. 73) y según algunos autores simplemente obsoleto (Bates 2002). No obstante, puesto que, hasta donde sabe el autor, no es posible la obtención automática de categorías para un sistema alternativo como la clasificación por facetas[2] y dado que la mayor parte de técnicas de categorización automática no suelen aprovechar tal estructura jerárquica[3] este capítulo se centrará en la categorización automática de documentos en categorías "planas", sin ningún tipo de relación jerárquica entre las mismas.

H. Information Systems
H.3 INFORMATION STORAGE AND RETRIEVAL
H.3.3 Information Search and Retrieval
Clustering
...
I. Computing Methodologies
I.5 PATTERN RECOGNITION
I.5.3 **Clustering**

"The cascade of interactions in the digital library interface"
(Marcia J. Bates 2002)

H. Information Systems
H.3 INFORMATION STORAGE AND RETRIEVAL
H.3.3 Information Search and Retrieval
H.3.5 On-line Information Services
H.3.7 Digital Libraries
H.5 INFORMATION INTERFACES AND PRESENTATION (I.7)
H.5.2 User Interfaces (D.2.2, H.1.2, I.3.6)

Fig. 73 Fragmento del sistema de clasificación de la *ACM* y categorización de un artículo.

A la izquierda se muestra un fragmento del sistema de clasificación de la ACM, nótese cómo una de las categorías (*clustering*) aparece como subcategoría de dos categorías distintas. A la derecha se muestra la clasificación de un artículo sobre interfaces de usuario de bibliotecas digitales; además de requerirse cuatro categorías para etiquetar dicho trabajo varias de las categorías están cruzadas (hacia I.7 Procesamiento de Documentos y Texto, D.2.2 Herramientas y Técnicas de Diseño, H.1.2 Sistemas Persona/Máquina e I.3.6 Metodología y Técnicas).

Las categorizaciones antes mencionadas fueron elaboradas por expertos humanos; sin embargo, la categorización automática de documentos no sólo es posible sino que constituye una herramienta muy útil para enfrentarse a la consabida sobrecarga de información.

Una aplicación inmediata es la de asignar temas a los documentos o *topic tagging*. Por ejemplo, Maarek y Ben Shaul (1996) emplearon esta técnica para organizar listas de enlaces favoritos (*bookmarks*), Hearst y Karadi (1997) para etiquetar grupos de documentos médicos extrayendo las categorías más frecuentes en cada conjunto y Attardi, Gullì y Sebastiani (1999) para etiquetar documentos web.

---

[1] Steinberg (1996, p. 3) relata una anécdota sobre la categorización de un sitio web en el directorio *Yahoo!* El sitio web de la *Messianic Jewish Alliance of America* (Alianza Judía Mesiánica de América) fue inicialmente adscrito a la categoría "Judaismo". Este hecho provocó quejas de ciertas organizaciones judías puesto que consideran a los judíos mesiánicos unos herejes ya que creen que Jesús es el mesías. Ante las protestas el sitio fue trasladado a la categoría "Cristiandad" causando entonces las quejas de la citada organización puesto que no son cristianos. Finalmente la página fue asignada a una categoría propia, "Judaismo Mesiánico", que actualmente contiene 220 enlaces.

[2] Sistematizada por S.R. Ranganathan (Chan 1994, p. 390).

[3] No obstante, se ha tratado de explotar la estructura jerárquica para mejorar el rendimiento de los categorizadores de documentos (Koller y Sahami 1997), (McCallum *et al.* 1998) o (Weigend, Wiener y Pedersen 1999)

No obstante, en estos trabajos no puede hablarse propiamente de categorización automática. En el primer y tercer caso no hay categorías predefinidas ya que las etiquetas se extraen automáticamente, mientras que en el segundo, aunque se utilizaban las categorías *MeSH*, no se empleaban para categorizar documentos sino para etiquetar *clusters*.

Chekuri *et al.* (1997), en cambio, emplearon las categorías de la taxonomía de *Yahoo!* para etiquetar sitios web y Li *et al.* (1999) al igual que Maarek y Ben Shaul desarrollaron un sistema para organizar *bookmarks* pero, al contrario que ellos, no utilizaron etiquetas automáticas sino las definidas en otras taxonomías (como por ejemplo la empleada por la Biblioteca del Congreso de EE.UU).

También se han aplicado técnicas de categorización automática para clasificar correo electrónico en categorías establecidas por el usuario (Cohen 1996), filtrar correo no deseado o *spam* (Sahami *et al.* 1998) o determinar si un documento es o no relevante para un usuario dado (Schütze, Hull y Pedersen 1995).

2 La categorización como un problema de aprendizaje automático

Las aplicaciones anteriormente descritas se enmarcan dentro de los problemas de aprendizaje supervisado que tiene como objetivo la obtención de una función a partir de datos de entrenamiento. En el caso de la categorización automática de documentos los datos de entrada serán vectores construidos a partir de los documentos y los valores de salida serán discretos (la categoría[1] a la que se asigna cada documento).

Ni que decir tiene que existen muy diversas técnicas para obtener dicha función y que se han aplicado muchas de ellas a la categorización de texto natural. Fabrizio Sebastiani (2002) publicó un magnífico estudio sobre el tema revisando aspectos fundamentales como la forma de representar los documentos, las distintas técnicas para construir categorizadores así como la manera de evaluar dichos categorizadores. No obstante, al igual que en capítulos anteriores, se describirán brevemente algunas de las técnicas más comunes.

Una forma sencilla de categorizar documentos consiste en examinar el texto en busca de algún término que permita discriminarlo. El correo electrónico constituye un magnífico ejemplo puesto que existen dos categorías principales: mensajes deseados y no deseados. Así, aquellos que incluyan términos como `congratulations`, `won`, `free`, `unlimited` o `mortgage` pueden desecharse con relativa confianza. Un usuario podría especificar reglas en su cliente de correo para eliminar este tipo de mensajes. Sin embargo, poco después tendría que añadir reglas para variaciones "tipográficas" de los mismos términos como `mortage` (*sic*) o `un1imited`. Por tanto, lo ideal sería emplear técnicas que, a partir de una serie de ejemplos, generasen las reglas apropiadas. Los árboles y reglas de decisión o, mejor dicho, los métodos para inducirlos son tales técnicas.

Un **árbol de decisión** constituye una agrupación jerárquica de reglas que permiten obtener un único valor objetivo a partir de un conjunto de datos de origen (véase Fig. 74). Las **reglas de decisión** son similares pero dan lugar a categorizadores más compactos. Por su propia naturaleza este tipo de técnicas son más intuitivas para el usuario final que podría, en caso necesario, añadir reglas propias. No obstante, aunque esto resulta una ventaja obvia los términos no suelen ponderarse sino que deben manejarse de forma binaria. Aun así, hay toda una serie de trabajos que han aplicado árboles –Lewis y Catlett (1994), Apté, Damerau y Weiss (1998) o Weiss *et al.* (1999)– o reglas de decisión –Apté, Damerau y Weiss (1994),

[1] Como se verá más adelante, en muchos problemas no resulta viable limitarse a una única categoría por cada documento y, así, a cada *ítem* se asociarán una o más categorías o etiquetas.

Cohen (1996), Moulinier y Ganascia (1996) o Li y Yamanishi (1999)– a la categorización automática de documentos y se han utilizado árboles de decisión como base de otros métodos de categorización como el *boosting* (que se verá más adelante).

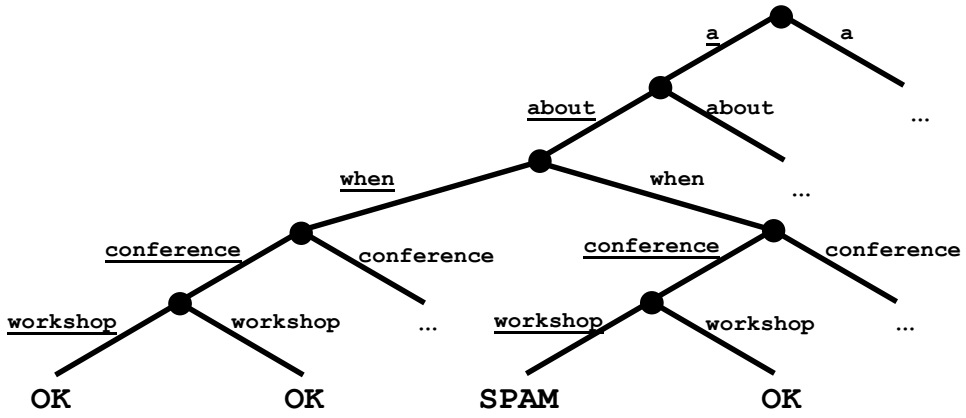


Fig. 74 Un árbol de decisión incompleto y muy sencillo para separar el *spam* recibido por un investigador hispanohablante.

Las ramas están etiquetadas con términos que pueden aparecer en los documentos. Los términos subrayados indican la ausencia del término y las hojas representan las distintas categorías.

Como se ha dicho, los árboles y reglas de decisión emplean una representación binaria de los documentos; sin embargo, es posible construir un categorizador de documentos basado en el modelo vectorial clásico de una forma muy sencilla. En la fase de "entrenamiento" simplemente se construye el vector para cada documento de la colección y se le asocia su categoría. En la fase de categorización se emplea el documento a clasificar como consulta y se obtienen k resultados, la clase más frecuente entre esos k documentos resultantes es la categoría a la que pertenece el nuevo documento.

Debido al trabajo prácticamente nulo que se realiza durante el entrenamiento este método es calificado como "perezoso" (Chakrabarti 2003, p. 134), por otro lado, requiere almacenar toda la información de entrenamiento y resulta computacionalmente costoso en el momento de la categorización. No obstante, puede refinarse y ofrecer resultados superiores a los obtenidos con árboles o reglas de decisión (Cohen y Hirsh 1998) (Han, Karypis y Kumar 1999) o próximos a los de técnicas más eficaces como *SVM* (Yang y Liu 1999). Una de las modificaciones más relevantes en esta técnica es el cambio del método de ponderación puesto que la técnica *tf*idf* aplicada sobre la colección de documentos de entrenamiento no es muy efectiva (Yang y Chute 1994, citado por Han *et al.* 1999) o (Chakrabarti 2003, p. 135).

Otro tipo de categorizadores muy populares son los denominados ***naïve Bayes*** (Minsky y Papert 1969, citado en Elkan 1997, p.3). Se trata de categorizadores probabilísticos basados en el teorema de Bayes y que reciben el apelativo *naïve* (simple) al suponer, de manera deliberada e irreal[1], una total independencia entre los valores que toman los distintos términos en cada clase. En la fase de entrenamiento este categorizador calcula

[1] El hecho de que los términos no son independientes entre sí resulta especialmente claro si se toma como ejemplo la poblada categoría de correo no solicitado; términos como `v1agra`, `anonymously`, `prescription`, `online` u `order` muestran una clara dependencia.

la probabilidad de aparición de cada término (normalmente palabras o *stems*) condicionada a cada categoría. En la fase de categorización se debe calcular la probabilidad de cada categoría condicionada a la frecuencia de aparición de los términos en el documento seleccionando la más probable.

A pesar de su simplicidad los categorizadores *naïve* Bayes muestran un buen rendimiento (Domingos y Pazzani 1997, p. 105-106) y, según Chakrabarti (2003, p. 175-176), tal vez su sencillez, su facilidad de implementación y la rapidez con que se adaptan a cambios en las colecciones de documentos explican su popularidad frente a otros algoritmos más efectivos. A pesar de todo, su aplicación a categorización de documentos es relativamente reciente: por ejemplo, Larkey y Croft (1996) los emplearon, individualmente y combinados con otros métodos, para asignar códigos *ICD*[1] a diagnósticos médicos, Koller y Sahami (1997) y Chakrabarti *et al.* (1998c) para categorizar documentos dentro de taxonomías y Sahami *et al.* (1998) para filtrar correo no deseado.

Otra técnica de aprendizaje automático que también se ha utilizado para categorizar documentos son las redes neuronales. Una **red neuronal** (McCulloch y Pitts 1943) (Rosenblatt 1958) es un sistema que conecta un conjunto de elementos de proceso muy simples en una serie de capas. La capa de entrada contiene tantos elementos como variables definan los ejemplares del problema y la capa de salida un elemento por cada categoría a detectar.

Schütze, Hull y Pedersen (1995), Wiener, Pedersen y Weigend (1995), Weigend, Wiener y Pedersen (1999), Ng, Goh y Low (1997) o Ruiz y Srinivasan (1997 y 1999) han aplicado con éxito redes neuronales a problemas de categorización de documentos. Yang y Liu (1999) mostraron que los resultados obtenidos con esta técnica son inferiores a los alcanzados con *k*-vecinos o *SVM* (que se verá más adelante) mientras que Lam y Lee (1999) analizaron en qué medida la reducción del número de términos puede mejorar el rendimiento de la red neuronal.

Ya se mencionaron en el capítulo anterior los mapas de Kohonen o mapas auto-organizativos (*SOM*) así como su relación con las redes neuronales. Roussinov y Chen (1998) y Ontrup y Ritter (2001) la han utilizado para categorizar documentos. Tan sólo los últimos ofrecen una comparativa con otra técnica, en este caso *k*-vecinos, afirmando que los resultados obtenidos por *SOM* se aproximan a los de dicho método.

Además del método de los *k*-vecinos hay otra técnica de categorización que no surgió del campo del aprendizaje automático sino del área de recuperación de información; se trata del **algoritmo de Rocchio** (1971) para expandir consultas por realimentación (*relevance feedback*). La idea es sencilla: (1) dada una consulta, un sistema de recuperación de información proporciona al usuario un conjunto de documentos, (2) el usuario selecciona los que considera relevantes y (3) se "enriquece" la consulta original calculando la diferencia entre los documentos relevantes (POS_i, véase la ecuación) y los no relevantes (NEG_i). Así, una categoría c_i estaría representada por un vector de pesos w_{ki} calculados según la siguiente fórmula en la que w_{kj} es el peso del término t_k en el documento d_j.

$$w_{ki} = \beta \cdot \sum_{\{d_j \in POS_i\}} \frac{w_{kj}}{|POS_i|} - \gamma \cdot \sum_{\{d_j \in NEG_i\}} \frac{w_{kj}}{|NEG_i|}$$

[1] *ICD – International Statistical Classification of Diseases and Related Health Problems* (Clasificación Estadística Internacional de Enfermedades y otros Problemas de la Salud) es un catálogo publicado por la *OMS* que describe de manera detallada enfermedades y heridas a las que asigna un código de hasta 5 caracteres.

Según Sebastiani (2002) fue Hull (1994) el primero en adaptar esta técnica, sin embargo, ese trabajo trata sobre un caso especial de categorización, la separación de una colección en conjuntos de documentos relevantes y no relevantes a partir de unos pocos ejemplares. A juicio del autor[1] sería más acertado atribuir el mérito de la adaptación a Ittner, Lewis y Ahn (1995) que emplearon Rocchio para asignar textos obtenidos a partir de imágenes de baja calidad a distintas categorías. Además de estos, otros investigadores han empleado o mejorado el algoritmo de Rocchio en este campo, por ejemplo Joachims (1997), Singhal, Mitra y Buckley (1997) o Schapire, Singer y Singhal (1998).

Spam-1	Powerful enlargement.
Spam-2	Find out about cialis. Viagra's big brother.
Spam-3	Viagra: save more buying more stylus sad.
Spam-4	Get viagra anonymously! Fast shipping.
Spam-5	Viagra, vallium, cialis.
Ham-1	Call for participation: constraints in discourse.
Ham-2	Call for participation and studentship applications for CLIMA VI.
Ham-3	Fast SVM training on very large data sets.
Ham-4	Job opening for nationals of EU enlargement countries.
Ham-5	Call for papers ICCBSS 2006.

Fig. 75 Una colección de 10 documentos y 2 categorías: *spam* (correo no deseado) y *ham* (correo deseado).

Para aclarar el funcionamiento del método de Rocchio se describirá un pequeño ejemplo de categorización de correo electrónico. En Fig. 75 se muestra la colección que se empleará para el "entrenamiento" y que consta de 10 breves documentos que pertenecen las categorías de correo deseado (*ham*) y no deseado (*spam*).

Descontando las palabras vacías se obtienen 34 términos distintos[2] para los que se calcula su valor *idf*. Posteriormente, se determina para cada documento su representación vectorial otorgando a cada término del documento su peso *tf*idf*. No obstante, puesto que no hay términos repetidos en ninguno de los textos, este peso resulta idéntico al valor *idf* original (véase Fig. 76).

Como se puede apreciar en la ecuación anterior el método de Rocchio es parametrizable mediante los valores β y γ que, básicamente, permiten controlar qué tipo de ejemplos (positivos o negativos) influyen más a la hora de construir la correspondiente categoría. En Fig. 77 y Fig. 79 se muestra el resultado de emplear $\beta=\gamma=1$ (ejemplos positivos y negativos tienen la misma influencia) y en Fig. 80 para $\beta=0{,}5$ y $\gamma=1$ (los ejemplos negativos tienen más influencia que los positivos).

Con independencia de los valores otorgados a ambos parámetros es necesario calcular el peso w_{ki} para cada término k de cada categoría i de acuerdo a la ecuación anterior (véase Fig. 77 y Fig. 80). Una vez calculado el módulo del vector de cada categoría ya ha finalizado el "entrenamiento" del categorizador.

Para llevar a cabo la categorización de un nuevo documento bastaría con comparar el vector correspondiente con los de las distintas categorías. Aquella categoría más próxima sería aquella a la que habría que asignar al documento. A fin de mostrar este funcionamiento se utilizarán los documentos que aparecen en Fig. 78.

[1] Juicio compartido por Cohen y Singer (1999, p. 155).

[2] Los términos se representan con el subíndice k en la ecuación.

Término	IDF	Spam-1	Spam-2	Spam-3	Spam-4	Spam-5	Ham-1	Ham-2	Ham-3	Ham-4	Ham-5
2006	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**
anonymously	1,00	0,00	0,00	0,00	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00
applications	1,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00	0,00
big	1,00	0,00	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
brother	1,00	0,00	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
buying	1,00	0,00	0,00	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00	0,00
cialis	0,50	0,00	**0,50**	0,00	0,00	**0,50**	0,00	0,00	0,00	0,00	0,00
clima	1,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00	0,00
constraints	1,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00	0,00	0,00
countries	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00
data	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00
discourse	1,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00	0,00	0,00
enlargement	0,50	**0,50**	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**0,50**	0,00
eu	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00
fast	0,50	0,00	0,00	0,00	**0,50**	0,00	0,00	0,00	0,50	0,00	0,00
iccbss	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**
job	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00
large	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00
nationals	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00
opening	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00
papers	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**
participation	0,50	0,00	0,00	0,00	0,00	0,00	**0,50**	**0,50**	0,00	0,00	0,00
powerful	1,00	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
sad	1,00	0,00	0,00	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00	0,00
save	1,00	0,00	0,00	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00	0,00
sets	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00
shipping	1,00	0,00	0,00	0,00	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00
studentship	1,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00	0,00
stylus	1,00	0,00	0,00	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00	0,00
svm	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00
training	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00
vallium	1,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00	0,00	0,00	0,00
vi	1,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00	0,00
viagra	0,25	0,00	**0,25**	**0,25**	**0,25**	**0,25**	0,00	0,00	0,00	0,00	0,00

Fig. 76 La colección anterior representada mediante el modelo vectorial.

Se han eliminado las palabras vacías y se ha simplificado el valor de *idf* como el inverso del número de documentos que contienen un término dado.

Término	Spam ✓	Spam ✖	WkSpam (β=γ=1)	Término	Ham ✓	Ham ✖	WkHam (β=γ=1)
2006	0,00	0,20	-0,20	2006	0,20	0,00	0,20
anonymously	0,20	0,00	0,20	anonymously	0,00	0,20	-0,20
applications	0,00	0,20	-0,20	applications	0,20	0,00	0,20
big	0,20	0,00	0,20	big	0,00	0,20	-0,20
brother	0,20	0,00	0,20	brother	0,00	0,20	-0,20
buying	0,20	0,00	0,20	buying	0,00	0,20	-0,20
cialis	0,20	0,00	0,20	cialis	0,00	0,20	-0,20
clima	0,00	0,20	-0,20	clima	0,20	0,00	0,20
constraints	0,00	0,20	-0,20	constraints	0,20	0,00	0,20
countries	0,00	0,20	-0,20	countries	0,20	0,00	0,20
data	0,00	0,20	-0,20	data	0,20	0,00	0,20
discourse	0,00	0,20	-0,20	discourse	0,20	0,00	0,20
enlargement	0,10	0,10	0,00	enlargement	0,10	0,10	0,00
eu	0,00	0,20	-0,20	eu	0,20	0,00	0,20
fast	0,10	0,10	0,00	fast	0,10	0,10	0,00
iccbss	0,00	0,20	-0,20	iccbss	0,20	0,00	0,20
job	0,00	0,20	-0,20	job	0,20	0,00	0,20
large	0,00	0,20	-0,20	large	0,20	0,00	0,20
nationals	0,00	0,20	-0,20	nationals	0,20	0,00	0,20
opening	0,00	0,20	-0,20	opening	0,20	0,00	0,20
papers	0,00	0,20	-0,20	papers	0,20	0,00	0,20
participation	0,00	0,20	-0,20	participation	0,20	0,00	0,20
powerful	0,20	0,00	0,20	powerful	0,00	0,20	-0,20
sad	0,20	0,00	0,20	sad	0,00	0,20	-0,20
save	0,20	0,00	0,20	save	0,00	0,20	-0,20
sets	0,00	0,20	-0,20	sets	0,20	0,00	0,20
shipping	0,20	0,00	0,20	shipping	0,00	0,20	-0,20
studentship	0,00	0,20	-0,20	studentship	0,20	0,00	0,20
stylus	0,20	0,00	0,20	stylus	0,00	0,20	-0,20
svm	0,00	0,20	-0,20	svm	0,20	0,00	0,20
training	0,00	0,20	-0,20	training	0,20	0,00	0,20
vallium	0,20	0,00	0,20	vallium	0,00	0,20	-0,20
vi	0,00	0,20	-0,20	vi	0,20	0,00	0,20
viagra	0,20	0,00	0,20	viagra	0,00	0,20	-0,20
		|spam|	**1,13**			**|ham|**	**1,13**

Fig. 77 Cálculo de los vectores para las categorías *spam* y *ham* (β=γ=1).

Empleando los mismos valores de β y γ (véase la ecuación en página 111) se "valoran" en la misma medida los ejemplos positivos y negativos para calcular el vector que representa a cada categoría.

La comparación entre los vectores de documento y categorías puede realizarse con cualquier medida pero en general se emplea la función del coseno (véase Fig. 79 y Fig. 80).

Los métodos descritos hasta el momento (p.ej. árboles y reglas de decisión, categorizadores bayesianos o Rocchio) emplean un único agente que "aprende" a partir de un único conjunto de ejemplos. Sin embargo, es posible emplear varios categorizadores y constituir lo que se denomina un **"comité"** de tal manera que cada miembro del mismo emita un voto para cada documento a categorizar. Una técnica distinta de los comités pero que también requiere la participación de varios categorizadores es la denominada ***boosting***[1]. La aproximación intuitiva es la siguiente: (1) un categorizador "aprende" sobre una parte del conjunto de entrenamiento y es probado sobre el resto del conjunto[2]; (2) aquellos documentos del conjunto de entrenamiento que clasifique mal, junto con algunos otros de su subconjunto de entrenamiento original, se utilizan para entrenar otro categorizador que, de este modo, "aprende" casos más "difíciles"; (3) el esquema se repite *n* veces.

spamTest Cheapest viagra, cialis delivered anonymously.
hamTest EU must defer enlargement if French vote no.

Término	spamTest	hamTest
2006	0,00	0,00
anonymously	**1,00**	0,00
applications	0,00	0,00
big	0,00	0,00
brother	0,00	0,00
buying	0,00	0,00
cialis	**0,50**	0,00
clima	0,00	0,00
constraints	0,00	0,00
countries	0,00	0,00
data	0,00	0,00
discourse	0,00	0,00
enlargement	0,00	**0,50**
eu	0,00	**1,00**
fast	0,00	0,00
iccbss	0,00	0,00
job	0,00	0,00
large	0,00	0,00
nationals	0,00	0,00
opening	0,00	0,00
papers	0,00	0,00
participation	0,00	0,00
powerful	0,00	0,00
sad	0,00	0,00
save	0,00	0,00
sets	0,00	0,00
shipping	0,00	0,00
studentship	0,00	0,00
stylus	0,00	0,00
svm	0,00	0,00
training	0,00	0,00
vallium	0,00	0,00
vi	0,00	0,00
viagra	**0,25**	0,00

Fig. 78 Un par de documentos de prueba y su traducción al modelo vectorial.

Naturalmente, es necesario garantizar que semejante método produce sistematicamente un aprendizaje efectivo. Esa garantía fue obtenida por Schapire a partir del trabajo de Valiant y Kearns. Valiant (1984) introdujo el modelo de aprendizaje ***PAC*** (*Probably Aproximately Correct*) en el cual el categorizador recibe ejemplos de una clase tomados al azar y debe producir una regla que permita discriminar nuevos ejemplares. Un categorizador "eficiente" encuentra una regla correcta, con una alta probabilidad, para todos los ejemplares excepto para una fracción establecida de manera arbitraria. Kearns (1988) introdujo el concepto de categorizador "débil", aquel cuya regla de decisión es sólo ligeramente mejor que una decisión al azar, y planteó la posibilidad de alcanzar un categorizador eficiente partiendo de categorizadores débiles (la denominada ***boosting hypothesis***).

[1] Literalmente "aumento", "promoción", "elevación", "empuje".

[2] Recuérdese en las colecciones empleadas para el entrenamiento y prueba de métodos de categorización los documentos deben estar "etiquetados" con su correspondiente categoría (o categorías).

WkSpamTest·WkSpam	WkSpamTest·WkHam	WkHamTest·WkSpam	WkHamTest·WkHam
0,00	0,00	0,00	0,00
0,20	-0,20	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,10	-0,10	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	-0,20	0,20
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00
0,05	-0,05	0,00	0,00

spamTest·spam	spamTest·ham	hamTest·spam	hamTest·ham
0,35	-0,35	-0,20	0,20

cosDis (spamTest, spam) $\beta=\gamma=1$	cosDis (spamTest, ham) $\beta=\gamma=1$	cosDis (hamTest, spam) $\beta=\gamma=1$	cosDis (hamTest, ham) $\beta=\gamma=1$
0,27	-0,27	-0,16	**0,16**

Fig. 79 Cálculo de la función del coseno entre los documentos de prueba y las categorías ($\beta=\gamma=1$).

Si se admiten vectores con pesos negativos (como es el caso) la función del coseno toma valores entre -1 (180º) y +1 (0º). Así, a mayor valor mayor similitud y en este caso la categoría más próxima será aquella a la que se asigne el documento. En el ejemplo el documento *spamTest* ha sido etiquetado correctamente como *spam* y *hamTest* como correo deseado.

Término	Spam ✓	Spam ✖	WkSpam (β=0.5, γ=1)
2006	0,00	0,20	-0,20
anonymously	0,20	0,00	0,10
applications	0,00	0,20	-0,20
big	0,20	0,00	0,10
brother	0,20	0,00	0,10
buying	0,20	0,00	0,10
cialis	0,20	0,00	0,10
clima	0,00	0,20	-0,20
constraints	0,00	0,20	-0,20
countries	0,00	0,20	-0,20
data	0,00	0,20	-0,20
discourse	0,00	0,20	-0,20
enlargement	0,10	0,10	-0,05
eu	0,00	0,20	-0,20
fast	0,10	0,10	-0,05
iccbss	0,00	0,20	-0,20
job	0,00	0,20	-0,20
large	0,00	0,20	-0,20
nationals	0,00	0,20	-0,20
opening	0,00	0,20	-0,20
papers	0,00	0,20	-0,20
participation	0,00	0,20	-0,20
powerful	0,20	0,00	0,10
sad	0,20	0,00	0,10
save	0,20	0,00	0,10
sets	0,00	0,20	-0,20
shipping	0,20	0,00	0,10
studentship	0,00	0,20	-0,20
stylus	0,20	0,00	0,10
svm	0,00	0,20	-0,20
training	0,00	0,20	-0,20
vallium	0,20	0,00	0,10
vi	0,00	0,20	-0,20
viagra	0,20	0,00	0,10
		|spam|	**0,96**

cosDis (spamTest, spam) β=0.5, γ=1	cosDis (spamTest, ham) β=0.5, γ=1
0,22	-0,44

Término	Ham ✓	Ham ✖	WkHam (β=0.5, γ=1)
2006	0,20	0,00	0,10
anonymously	0,00	0,20	-0,20
applications	0,20	0,00	0,10
big	0,00	0,20	-0,20
brother	0,00	0,20	-0,20
buying	0,00	0,20	-0,20
cialis	0,00	0,20	-0,20
clima	0,20	0,00	0,10
constraints	0,20	0,00	0,10
countries	0,20	0,00	0,10
data	0,20	0,00	0,10
discourse	0,20	0,00	0,10
enlargement	0,10	0,10	-0,05
eu	0,20	0,00	0,10
fast	0,10	0,10	-0,05
iccbss	0,20	0,00	0,10
job	0,20	0,00	0,10
large	0,20	0,00	0,10
nationals	0,20	0,00	0,10
opening	0,20	0,00	0,10
papers	0,20	0,00	0,10
participation	0,20	0,00	0,10
powerful	0,00	0,20	-0,20
sad	0,00	0,20	-0,20
save	0,00	0,20	-0,20
sets	0,20	0,00	0,10
shipping	0,00	0,20	-0,20
studentship	0,20	0,00	0,10
stylus	0,00	0,20	-0,20
svm	0,20	0,00	0,10
training	0,20	0,00	0,10
vallium	0,00	0,20	-0,20
vi	0,20	0,00	0,10
viagra	0,00	0,20	-0,20
		|ham|	**0,83**

cosDis (hamTest, spam) β=0.5, γ=1	cosDis (hamTest, ham) β=0.5, γ=1
-0,28	**0,09**

Fig. 80 Cálculo de los vectores para las categorías *spam* y *ham* y de la similitud entre documentos de prueba y categorías (β=0.5 y γ=1).

Los valores β y γ utilizados en este caso dan mayor importancia a los ejemplos negativos que a los positivos. Nuevamente la categorización de los documentos de prueba es correcta.

Schapire (1990) demostró la equivalentecia entre ambos tipos de aprendizaje y que, efectivamente, resultaba factible construir un categorizador eficiente (en el marco del aprendizaje *PAC*) partiendo de categorizadores débiles. El método de *boosting*, por tanto, emplea un algoritmo de categorización cualquiera y construye de manera secuencial una serie de categorizadores que tratan de mejorar los peores resultados obtenidos por el categorizador anterior. El propio Schapire es el responsable de la mayor parte de las aplicaciones del *boosting* a la categorización de documentos, por ejemplo, Schapire *et al.* (1998) o Schapire y Singer (2000).

Por último se describirá la técnica que ofrece los mejores resultados (Joachims 1998) (Dumais *et al.* 1998) (Kwok 1998) (Yang y Liu 1999), la conocida como ***Support Vector Machines*** o ***SVM***. Este método fue propuesto originalmente por Boser, Guyon y Vapnik (1992) y consiste en la transformación de los vectores de entrada, que definen una dimensión en la cual no son linealmente separables, a una dimensión superior que permita su separación mediante una única (hiper)superficie (véase Fig. 81).

La aplicación de la técnica *SVM* a categorización de texto se debe[1] a Joachims (1998) aunque tanto Dumais *et al.* (1998) como Kwok (1998) hicieron de forma casi simultánea la misma propuesta. A partir de ese momento otros investigadores han empleado

[1] Según Sebastiani (1999, p. 36).

SVM para categorizar documentos: Drucker, Wu y Vapnik (1999), Schohn y Cohn (2000), Tong y Koller (2000), Rennie y Rifkin (2001) o Jalam y Teytaud (2001).

Ya se ha mencionado que los categorizadores *naïve* Bayes son muy utilizados aun cuando el método *SVM* es superior. Tal vez la complejidad del método de entrenamiento, cuadrática, pudiese afectar negativamente a su popularización al precisar de conjuntos de entrenamiento relativamente pequeños y, en consecuencia, limitados. No obstante, Osuna, Freund y Girosi (1997), Kaufman (1998, citado en Platt 1999), Platt (1999), Cauwenberghs y Poggio (2000) o Collobert, Bengio y Bengio (2002) plantearon diversas formas de solucionar ese inconveniente acelerando la fase de entrenamiento y/o facilitando la manipulación de colecciones cambiantes. No obstante, la razón fundamental probablemente sea la "implementación intimidante" (Platt 1998) de las *SVMs* frente a métodos menos efectivos pero mucho más sencillos.

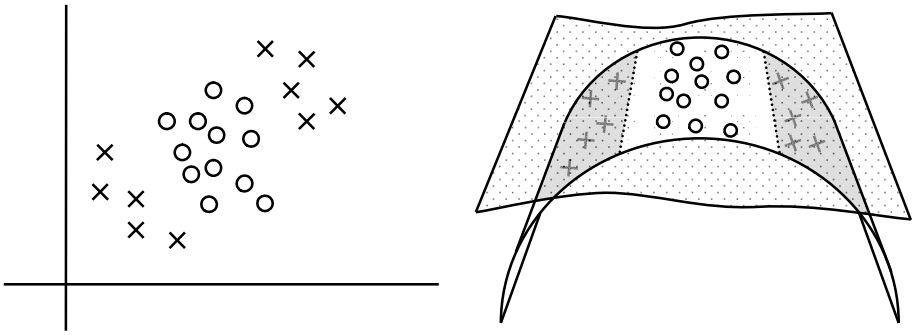


Fig. 81 Aproximación intuitiva al método de vectores soporte.

Los datos del problema original definen un espacio bidimensional donde no son linealmente separables. Transformandolos en un espacio tridimensional ya son separables mediante un plano.

3 Categorización de documentos con *blindLight*

Para categorizar una serie de documentos utilizando *blindLight* tanto los documentos como las categorías deben estar disponibles en la forma de vectores de *n*-gramas tal y como se describieron en los capítulos anteriores. Para obtener un vector de una categoría es posible emplear un único documento de muestra[1] aunque es más habitual realizar un entrenamiento sobre un conjunto de ejemplos. En ese caso el procedimiento es muy simple: (1) calcular un vector de *n*-gramas para cada documento de entrenamiento, (2) calcular el centroide de cada categoría, (3) calcular el centroide de todos los documentos de entrenamiento y (4) restar al centroide de cada categoría el centroide del conjunto de entrenamiento.

Una vez terminada la fase de aprendizaje se dispone de un único vector por categoría, vector susceptible de ser comparado con los vectores correspondientes a los documentos a categorizar. Aunque podría utilizarse la medida anteriormente descrita *PiRo* (véase página 66) ésta no proporciona resultados excesivamente satisfactorios. La razón es simple: dado que el número de *n*-gramas distintos aumenta[2] con la cantidad de documentos

[1] En el siguiente apartado se presentará un categorizador basado en *blindLight* que emplea un único documento para representar cada categoría.

[2] Crowder y Nicholas (1996) muestran cómo a medida que crece el número de documentos también aumenta el número de *n*-gramas distintos hasta alcanzar una meseta.

procesados los vectores de las distintas categorías y de los documentos a categorizar tienen tamaños y significatividades muy dispares. La diferencia en el tamaño es irrelevante pero una diferencia de significatividad notoria puede llevar a que las medidas Π y P tengan órdenes de magnitud diferentes y, por tanto, no tenga sentido combinarlas en una única medida de similitud.

Para evitar este problema se propone la siguiente versión "normalizada" de *PiRo*. Π es el cociente de la significatividad total de la intersección de documento y categoría entre la significatividad total del documento, P es el cociente de la significatividad total de la intersección entre la significatividad total de la categoría y *n* y *m* son, respectivamente, el número de *n*-gramas en los vectores categoría y documento. En todos los experimentos que se describen a lo largo del capítulo se ha utilizado esta medida con $\alpha=1-\alpha=0{,}5$[1].

$$PiRoNorm = \alpha \cdot \Pi + (1-\alpha) \cdot \frac{n}{m} \cdot \mathrm{P}$$

Por último, en la fase de categorización para cada documento incógnita se obtiene el vector de *n*-gramas correspondiente, y se calculan los valores de Π y P, y en consecuencia de *PiRoNorm*, para todas las categorías disponibles. En caso de tratarse de un problema de categorización en el que sólo deba asignarse una etiqueta al documento[2] se toma aquella categoría que obtenga para ese documento el valor máximo de *PiRoNorm*. En aquellos problemas en que puedan asignarse varias etiquetas a cada documento[3] tan sólo se proporciona una lista ordenada de todas las categorías y, por el momento, no se hace intento alguno por limitar el número de etiquetas asignadas.

4 Identificación automática del idioma a partir de un texto

Identificar el idioma en que está escrito un texto constituye un caso particular dentro de la categorización de documentos y es una tarea habitualmente requerida por buscadores web o en colecciones multi-idioma. Se han implementado múltiples sistemas para realizar este trabajo (Beesley 1988), (Cavnar y Trenkle 1994), (Dunning 1994), (Grefenstette 1995), (Sibun y Reynar 1996) o (Prager 1999) y todos han alcanzado una precisión casi perfecta. El único aspecto mejorable era el tiempo de ejecución y recientemente se ha presentado una nueva técnica (Poutsma 2002) que, aun cuando ofrece una precisión ligeramente peor que la mejor técnica disponible, es 85 veces más rápida que ésta.

En este sentido, la utilización de *blindLight* para identificar distintos idiomas sería una aportación poco significativa pues resulta muy difícil mejorar sustancialmente la precisión. No obstante, sí sería interesante determinar el rendimiento de esta nueva técnica en dos situaciones habituales al procesar información textual en Internet. A saber,

---

[1] El carácter *ad hoc* de esta medida no resulta totalmente satisfactorio. En la página 141 se muestran otras medidas de similitud y se discute la posibilidad de emplear programación genética para descubrir otras puesto que los valores Π y P son constantes para cada par de documentos comparados. No obstante, esto no invalida el hecho de que *blindLight* puede emplearse para llevar a cabo categorización automática, simplemente no se puede afirmar que esta medida sea la que ofrece una categorización óptima en todos los casos.

[2] Como por ejemplo en los experimentos descritos en los dos apartados siguientes: identificación de idioma y filtrado de correo no solicitado.

[3] Como en las colecciones `Reuters-21578` y `OHSUMED`.

documentos muy cortos (por ejemplo, el texto de consultas realizadas por los usuarios) y documentos con "ruido" (errores ortográficos, cabeceras de correo electrónico o *USENET*, etiquetas *HTML*, *Javascript*, etc).

A fin de comparar un identificador de idiomas basado en *blindLight* con otros métodos sería interesante disponer de alguna colección "estándar". Lamentablemente, la mayor parte de los autores elaboraron sus propias colecciones que no están disponibles; por su parte, Grefenstette y Sibun y Reynar sí emplearon una colección pública pero no libre[1].

Por suerte, algunos de los sistemas de identificación de idiomas están disponibles[2] *online*. Así, *TEXTCAT*[3] es una implementación del método de Cavnar y Trenkle capaz de identificar 70 idiomas distintos; *XEROX*[4] dispone de una herramienta aparentemente relacionada con los trabajos de Beesley y Grefenstette que soporta 47 idiomas y también existe una aplicación[5] de *Acquaintance* (Damashek 1995) mencionada en capítulos anteriores que distingue 66 idiomas y dialectos.

Según Poutsma (2002) el método que ofrece mayor precisión, incluso con muestras de unas decenas de caracteres, es el de Cavnar y Trenkle. En cuanto al sistema de *XEROX*, al ser de código cerrado (aunque se puede usar *online* de forma gratuita), no existe ninguna publicación reciente que describa la actual implementación ni analice su rendimiento de modo que el único "respaldo" a su eficiencia es el hecho de que se trata de un producto comercial. Así pues, al comparar *blindLight* con las tres herramientas citadas se estaría enfrentando a métodos cuya precisión ha sido sobradamente contrastada. Sin embargo, antes de describir los experimentos llevados a cabo es necesario comentar brevemente el modo en que las técnicas anteriores identifican los idiomas.

Tanto *TEXTCAT* como *Acquaintance* utilizan *n*-gramas de caracteres para construir el modelo de los lenguajes y representar las muestras de texto. La diferencia entre ambos radica en la manera en que se comparan muestra y modelo. Cavnar y Trenkle (1994) utilizan una medida denominada *out-of-place* ("fuera de lugar") que consiste básicamente en ordenar, basándose en su frecuencia de aparición, los *n*-gramas de muestra y modelo y determinar para cada *n*-grama de la muestra si ocupa el mismo puesto en el modelo o cuán desplazado se encuentra[6]. En el caso de *Acquaintance* se emplea el producto escalar.

Por lo que se refiere al sistema de *XEROX*, no es posible hacer ninguna afirmación categórica puesto que Grefenstette (1995) describe dos técnicas: una basada en el uso de trigramas de caracteres y otra basada en la utilización de palabras comunes[7]. No obstante, Langer (2001) describe un identificador híbrido (utilizado por el buscador *AllTheWeb*[8]) que primeramente trata de identificar el idioma empleando palabras comunes y sólo en caso de no obtener un resultado fiable recurre a los *n*-gramas. Por ello, resulta razonable suponer que el actual sistema de *XEROX* opere de manera similar empleando los dos métodos descritos por Grefenstette (1995) para identificar el idioma utilizado en un texto.

---

[1] *European Corpus Initiative CD-ROM*.

[2] Enero de 2005.

[3] `http://odur.let.rug.nl/~vannoord/TextCat/`

[4] `http://www.xrce.xerox.com/competencies/content-analysis/tools/guesser.en.html`

[5] `http://complingone.georgetown.edu/~langid/`

[6] Se trata de una medida similar en cierto modo al coeficiente de correlación de Spearman (pág. 75).

[7] Estas palabras comunes serían las "palabras vacías" mencionadas en otros capítulos.

[8] `http://www.alltheweb.com`

Es necesario señalar, además, que en ninguno de los casos existe información sobre la cantidad de texto que se ha empleado para construir el modelo de cada idioma. Es éste un dato especialmente relevante puesto que, según Dunning (1994), un sistema entrenado sobre 50 Kbytes de texto alcanza una precisión del 99,9% sobre muestras de 500 bytes frente al 97% de un sistema entrenado con sólo 5 Kbytes.

Una versión preliminar del identificador basado en *blindLight* se describe en (Gayo Avello *et al.* 2004b). Dicho sistema era capaz de identificar 14 idiomas[1] y se había "entrenado" a partir de documentos de unos 10 Kbytes, en todos los casos los tres primeros capítulos del libro del Génesis. Posteriormente se cambió el texto modelo, aunque el tamaño apenas cambió, y se incrementó el número de idiomas: se escogió la Declaración Universal de Derechos Humanos y los idiomas identificables pasaron a ser 37[2].

Por lo que respecta al método de categorización es muy simple. Para cada idioma se construye un vector de *n*-gramas a partir de los documentos modelo; al recibir una muestra de un texto desconocido se obtiene su correspondiente vector y éste es comparado mediante *PiRoNorm* (véase pág. 118) con los vectores anteriores. El idioma que muestre un mayor parecido será el asignado a la muestra.

Como se dijo antes se perseguían dos objetivos con los experimentos: por un lado, determinar el comportamiento de *blindLight* y otros identificadores frente a textos muy cortos y, por otro, comprobar la tolerancia al ruido de los distintos métodos.

Para llevar a cabo la primera prueba se tomaron los temas[3] utilizados en las campañas de 2003 y 2004 del *CLEF* (*Cross Language Evaluation Forum*) un ejemplo de los cuales aparece en Fig. 82. Los temas estaban disponibles en alemán, castellano, finés, francés e inglés para 2003 y 2004 además de en italiano y sueco para el último año. La campaña de 2003 ofreció 60 temas por idioma y la de 2004 50 temas. Para cada tema e idioma se elaboraron 7 documentos combinando los distintos elementos constituyentes[4] obteniendo de este modo 4.550 documentos que contenían desde una única palabra hasta cerca de 100.

Una vez elaborada esta colección cada uno de los sistemas identificadores fue aplicado sobre la misma anotando el idioma asignado a cada documento y comparando la identificación con el idioma en que estaba originalmente escrito[5]. Los resultados obtenidos se muestran en la Tabla 10 y la Tabla 11 y de ellos se deduce que la técnica propuesta por el autor es sustancialmente mejor que las de (Cavnar y Trenkle 1994) y (Damashek 1995) para textos de entre 1 y 5 palabras[6] pero no consigue superar al sistema de *XEROX* en esos

---

[1] Alemán, castellano, catalán, danés, faroés, finés, francés, holandés, inglés, italiano, noruego, portugués, sueco y vasco.

[2] A los anteriores se añadieron: asturiano, bretón, corso, croata, eslovaco, estonio, frisón, gaélico escocés, gaélico irlandés, galés, gallego, húngaro, islandés, latín, letón, lituano, maltés, occitano auvergnat, occitano languedoc, polaco, rumano, sardo y turco.

[3] En cada campaña *CLEF* se ofrece un conjunto de temas en varios idiomas para elaborar (de manera automática o manual) las consultas que se utilizarán para evaluar los sistemas *IR*. Estos temas tratan sobre un asunto específico y constan de tres partes: (1) un breve título, (2) una descripción más verbosa de la necesidad de información a satisfacer y (3) una serie de criterios que se emplearán para juzgar la relevancia de los documentos retornados por el sistema.

[4] Cada tema *CLEF* consta de título (T), descripción (D) y narración (N). Así, los documentos construidos para cada tema serían: T, D, N, TD, TN, DN y TDN.

[5] En algunos casos el documento no estaba escrito en el idioma de "consulta". Dos temas de *CLEF'04* tenían por títulos `Lady Diana` y `Christopher Reeve`.

[6] En realidad *blindLight* es sustancialmente superior a *TEXTCAT* para textos de hasta 20 palabras.

mismos documentos. Cabe pensar si el hecho de que este último emplee listas de palabras vacías (conocimiento lingüístico) como una de las "pistas" para identificar el idioma pueda influir en un rendimiento superior para los textos más cortos.

```
<top>
     <num>
          C154
     </num>
     <ES-title>
          Libertad de Expresión en Internet
     </ES-title>
     <ES-desc>
          Encontrar documentos en los que se hable sobre la
          censura y la libertad de expresión en Internet.
     </ES-desc>
     <ES-narr>
          Los documentos en los que se discutan asuntos
          como la pornografía o el racismo en Internet, sin
          mencionar el tema de la censura o libertad de
          expresión, no se considerarán relevantes.
     </ES-narr>
</top>
```

Fig. 82 Un tema para la campaña de 2003 del *CLEF* escrito en castellano.

Los temas del *CLEF* describen necesidades de información que deben ser resueltas por un sistema *IR* obteniendo documentos de una colección de artículos periodísticos. Cada tema incluye un título corto, una descripción (normalmente una oración que indica la consulta) y una descripción más larga de las cualidades deseables en los documentos relevantes y, ocasionalmente, características de documentos no relevantes.

Longitud del texto	*blindLight*	Xerox	TEXTCAT	*Acquaintance*
1 o 2 palabras	42,37%	54,62%	10,97%	31,22%
3 a 5 palabras	65,91%	79,24%	26,90%	56,85%
6 a 10 palabras	90,13%	94,57%	60,10%	84,55%
11 a 15 palabras	95,10%	98,04%	78,51%	94,54%
16 a 20 palabras	98,48%	99,67%	87,68%	97,71%
21 a 30 palabras	99,45%	99,80%	92,13%	98,99%
31 a 50 palabras	99,86%	100,00%	96,06%	99,92%
Más de 50 palabras	100,00%	100,00%	99,42%	99,88%

Tabla 10. Precisión (macropromediada, *macroaveraged*[1]) de los cuatro identificadores.

El identificador basado en *blindLight* es sustancialmente mejor que *Acquaintance* para textos de entre 1 y 20 palabras, apreciablemente mejor para textos de 21 a 30 y la diferencia es inapreciable para textos más largos. *blindLight* es sustancialmente mejor que *TEXTCAT* para textos de entre 1 y 5 palabras y las diferencias son inapreciables para textos más extensos. El identificador de *XEROX* supera a *blindLight* de manera sustancial al identificar textos de 1 a 5 palabras y ofrece resultados análogos con muestras mayores.

[1] David D. Lewis (1991) señala dos formas de obtener resultados promedio al evaluar un sistema de categorización a las que denomina *macro-* y *microaveraging*. Para un conjunto de D documentos y una serie de K categorías un categorizador toma $D \cdot K$ decisiones evaluables individualmente. A fin de obtener un valor promedio puede calcularse la media de las evaluaciones de las decisiones correspondientes a cada categoría (*macroaveraging*) o bien tomar las $D \cdot K$ decisiones en conjunto (*microaveraging*). La diferencia entre uno y otro método de evaluación es simple: en el caso de *microaveraging* tiene más influencia el resultado global (número total de categorizaciones correctas) frente a las diferencias entre categorías (puede haber diferencias notables entre los resultados obtenidos para cada categoría) mientras que en el caso de *macroaveraging* influyen más las diferencias entre categorías que los resultados tomados en su conjunto, es decir, se "premiaría" al categorizador que obtiene resultados similares en todas las categorías. Dependiendo de la aplicación debe decidirse qué tipo de resultados son preferibles y emplear un método u otro de promedio para evaluar las distintas técnicas. Si se considera que todos los idiomas son igualmente importantes a la hora de ser correctamente identificados entonces los identificadores deberían evaluarse empleando *macroaveraging*. En cambio, si se considera que esto no es así, bien por el número de hablantes o por la cantidad de documentos existentes, debería optarse por evaluar mediante *microaveraging*. Langer (2001) proporciona datos interesantes sobre la importancia relativa de distintos idiomas encontrados en el índice de *AllTheWeb*.

Longitud del texto	*blindLight*	Xerox	TEXTCAT	*Acquaintance*
1 o 2 palabras	49,03%	63,23%	37,42%	9,03%
3 a 5 palabras	64,61%	76,25%	55,34%	24,70%
6 a 10 palabras	91,72%	94,70%	88,41%	65,23%
11 a 15 palabras	96,53%	98,07%	95,18%	78,03%
16 a 20 palabras	98,84%	99,67%	97,19%	87,27%
21 a 30 palabras	99,56%	99,78%	98,90%	92,76%
31 a 50 palabras	99,82%	100,00%	99,91%	95,70%
Más de 50 palabras	100,00%	100,00%	99,81%	99,25%

Tabla 11. Precisión (micropromediada, *microaveraged*) de los cuatro identificadores.

Las diferencias entre identificadores al comparar los resultados micropromediados son similares a los encontradas al comparar los datos macropromediados.

Por lo que respecta al segundo experimento se empleó la colección `1500-5LNG`[1] elaborada por el propio autor (Gayo Avello *et al.* 2004b). Dicha colección consta de 1500 artículos publicados en los grupos `soc.culture.basque`, `catalan`, `french`, `galiza` y `german`, es decir, contiene documentos escritos, teóricamente, en vasco, catalán, francés y alemán. El objetivo era idéntico al de Cavnar y Trenkle (1994), obtener textos escritos presumiblemente en un idioma (por ejemplo, catalán en el caso de `soc.culture.catalan`) a fin de probar el identificador de manera sencilla.

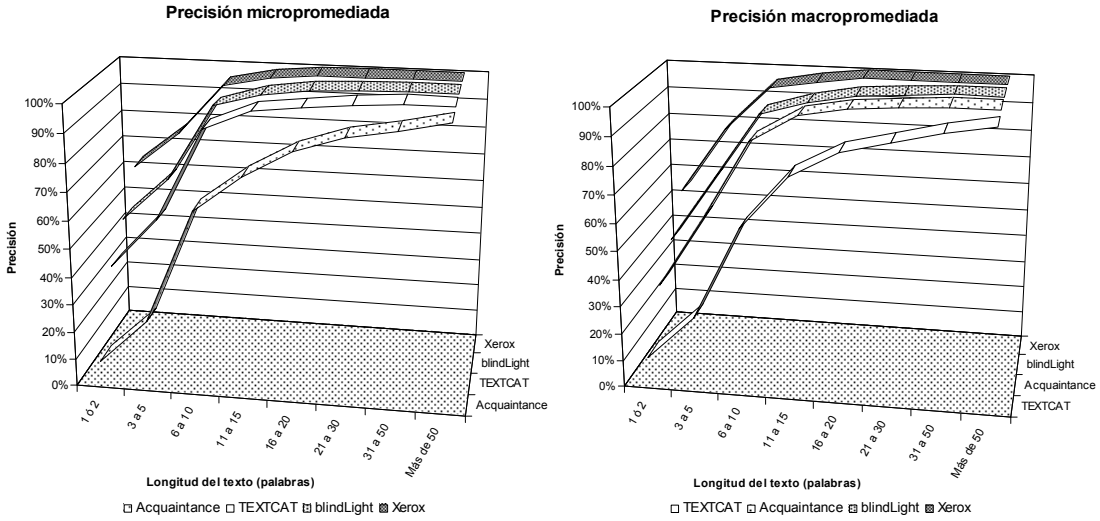


Fig. 83 Precisión de los identificadores en relación con la longitud del texto.

No obstante, estos grupos sufren de graves problemas de *spam* y publicación cruzada (*cross-posting*) por lo que los idiomas que aparecen en cada grupo son realmente diversos. Así, se emplean los siguientes idiomas: alemán, castellano, catalán, francés, gallego, inglés, italiano y vasco; mezclando en muchos artículos dos y, en ocasiones, tres idiomas. Por ello, y teniendo en cuenta que el sistema de *XEROX* no "conoce" el gallego, se eliminaron todos los artículos escritos en gallego (que no todos los artículos de `soc.culture.galiza`) además de aquellos en los que no había un predominio claro de un único idioma en el cuerpo del artículo. De este modo, la colección se redujo a 1358 documentos escritos en alemán, castellano, catalán, francés, inglés, italiano y vasco y que incluían las correspondientes cabeceras a modo de "ruido" (véase Fig. 84). Debido al escaso número de artículos en italiano y vasco (4 y 5, respectivamente) se evaluaron los distintos

[1] `http://www.di.uniovi.es/~dani/downloads/1500-5LNG.zip`

categorizadores con los restantes idiomas; los resultados obtenidos en este segundo experimento se muestran en Tabla 12 y Tabla 13.

```
From: unrien.dutout@nulle.part.fr (Unrien Dutout)
Newsgroups: soc.culture.french
Subject: C'est chouette ici...
Date: 4 Dec 2003 15:44:31 -0800
Organization: http://groups.google.com
Lines: 1
Message-ID: <67ec58c5.0312041544.218cc3d8@posting.google.com>
NNTP-Posting-Host: 82.66.227.13
Content-Type: text/plain; charset=ISO-8859-1
Content-Transfer-Enconding: 8bit
X-Trace: posting.google.com 1070581471 1210 127.0.0.1 (4 Dec 2003 23:44:31 GMT)
X-Complaints-To: groups-ab...@google.com
NNTP-Posting-Date: Thu, 4 Dec 2003 23:44:31 +0000 (UTC)

ça sent la déconfiture
```

Fig. 84 Un artículo escrito en francés con más del 90% de ruido.

Ruido en el texto	*blindLight*	Xerox	TEXTCAT	*Acquaintance*
0-5%	100,00%	99,24%	99,24%	98,47%
5-10%	100,00%	95,89%	95,43%	96,80%
10-15%	97,50%	97,50%	98,00%	98,50%
15-20%	96,32%	95,09%	95,71%	95,09%
20-25%	98,09%	95,54%	95,54%	98,73%
25-30%	96,49%	85,09%	91,23%	96,49%
30-35%	93,98%	77,11%	85,54%	98,80%
35-40%	86,57%	74,63%	82,09%	94,03%
40-50%	74,47%	63,83%	64,89%	81,91%
Más del 50%	73,33%	67,50%	47,50%	65,83%

Tabla 12. Precisión (micropromediada, *microaveraged*) de los cuatro identificadores.

Al micropromediar los resultados se comprueba que *blindLight* es sustancialmente superior al sistema de *XEROX* con más de un 25% de ruido; apreciablemente superior a *TEXTCAT* también a partir de ese mismo punto y sustancialmente a partir de un 40%. Al compararlo con *Acquaintance* las diferencias son inapreciables hasta un 30% de ruido, el rendimiento es peor entre un 30 y un 50% y sólo es materialmente superior con más de un 50% de ruido en el texto.

El análisis de estos resultados muestra que en términos absolutos *blindLight* es ligeramente superior al resto de técnicas cuanto mayor es la cantidad de ruido en el texto. No obstante, al macropromediar los resultados se comprueba que aunque, efectivamente, la técnica del autor es capaz de identificar correctamente un número mayor de documentos en presencia de ruido existen importantes diferencias de precisión entre los distintos idiomas[1] por lo que, tampoco en este experimento, se trata de la técnica más efectiva aunque los resultados son muy parejos hasta niveles de ruido del 30% y sólo *Acquaintance* se muestra claramente superior a *blindLight* a partir del 35%.

No obstante, como ya se dijo anteriormente parámetros importantes de los sistemas de referencia tales como la cantidad y la naturaleza del texto utilizada para construir los modelos de los otros sistemas de identificación son desconocidos por el autor. Sería interesante determinar si entrenando sobre otros tipos de documentos, empleando más texto[2] o utilizando otros tamaños de *n*-grama[3] mejoraría la precisión y en qué medida. Sin embargo, el autor considera que no es absolutamente necesario llegar en este trabajo a ese

[1] *blindLight* alcanzó una precisión de 100% y 94% en inglés y francés frente al 86,26% y 75,4% de castellano y alemán.

[2] Recuérdese que este identificador *blindLight* utilizó alrededor de 11 Kbytes por idioma

[3] El identificador *blindLight* aquí descrito empleó trigramas.

nivel de detalle y considera razonablemente argumentado que también en esta tarea la técnica que propone alcanza niveles de rendimiento comparables a los de métodos *ad hoc*.

Ruido en el texto	*blindLight*	Xerox	TEXTCAT	*Acquaintance*
0-5%	100,00%	99,17%	99,76%	99,52%
5-10%	100,00%	98,27%	97,96%	98,65%
10-15%	95,82%	98,94%	99,15%	99,36%
15-20%	95,22%	97,71%	98,25%	97,09%
20-25%	93,33%	98,08%	97,63%	98,89%
25-30%	94,49%	93,98%	95,60%	97,60%
30-35%	98,46%	90,26%	93,33%	99,23%
35-40%	79,50%	81,91%	88,21%	92,19%
40-50%	73,57%	77,86%	73,49%	83,55%
Más del 50%	54,67%	65,38%	44,38%	64,20%

Tabla 13. Precisión (macropromediada, *macroaveraged*) de los cuatro identificadores.

Los resultados macropromediados señalan que hasta un 30% de ruido los cuatro sistemas son análogos. Entre un 30 y un 35% de ruido la técnica del autor es apreciablemente mejor que la de *XEROX* y Cavnar y Trenkle (1994) y similar a *Acquaintance*. Con más de un 35% de ruido *blindLight* se comporta de manera peor que *XEROX* y *Acquaintance* aunque sólo la última es sustancialmente superior y sólo supera a *TEXTCAT* de manera sustancial con más de un 50% de ruido en el texto.

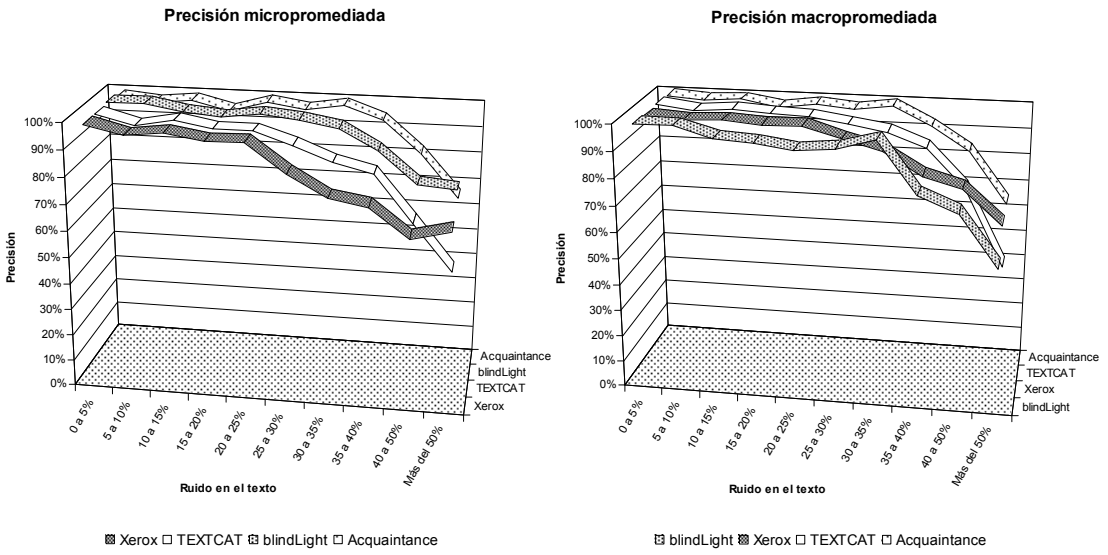


Fig. 85 Precisión de los identificadores en relación con el porcentaje de ruido.

5 Identificación de la autoría de un documento

Otro caso particular de categorización de documentos es la "atribución automática de autoría", es decir, la identificación del autor de un documento basándose en otros textos de dicho autor. Es necesario decir que no se considera que esta técnica sea especialmente fiable; Rudman (1998, p. 351), por ejemplo, afirma:

> *Los estudios no tradicionales de atribución de autoría –aquellos que emplean el ordenador, la estadística y la estilística– han tenido tiempo suficiente para superar cualquier "período transitorio" y entrar en una fase marcada por estudios sólidos, científicos y en constante progreso. Sin embargo, después de 30 años y 300 publicaciones no lo han hecho.*

En ese trabajo Rudman hace un repaso bastante exhaustivo tanto de las técnicas habituales como de las principales críticas a las mismas y a sus resultados. Esta situación es

análoga a la pugna existente entre los métodos tradicionales y estadísticos para clasificar lenguajes naturales y, como ésta, deberá ser resuelta por los investigadores involucrados en el campo. El autor tan sólo se ha aproximado a un problema clásico y bien estudiado dentro de la atribución de autoría: los denominados *Federalist Papers* (Artículos Federalistas).

Los *Federalist Papers* son una serie de 85 artículos publicados durante 1787 y 1788 en distintos periódicos del estado de Nueva York para convencer a los votantes de dicho estado sobre la necesidad de ratificar la futura constitución de los EE.UU. Dichos artículos aparecieron bajo el pseudónimo de Publius y fueron escritos por Alexander Hamilton, James Madison y John Jay. Posteriormente se llegó a un consenso sobre la autoría de cada artículo a excepción de 12, sobre los cuales sólo se estaba de acuerdo en que eran de Hamilton o de Madison.

Estos doce artículos son los conocidos como *disputed Federalist Papers* (los Artículos Federalistas disputados) y han generado bastante bibliografía: Mosteller y Wallace (1964, citado por Fung 2003) concluyeron, por métodos estadísticos, que los doce artículos en disputa eran obra de Madison. Desde entonces otros investigadores han empleado diversas técnicas[1] para alcanzar la misma conclusión y el autor también utilizó dicha colección como campo de prueba de la técnica *blindLight*.

Es preciso señalar que no se pretende hacer ninguna afirmación sobre la calidad de la técnica propuesta en el específico campo de la atribución de autoría puesto que Stamatos, Fakotakis y Kokkinakis (2001) advierten sobre las diferencias existentes entre los *Federalist Papers* y los textos que habitualmente se manejan en problemas de identificación de autoría. Simplemente se toma el problema de los artículos disputados como una tarea interesante de categorización.

El texto de los *Federalist Papers* fue obtenido en la Web[2] y procesado con *blindLight* empleando trigramas. Es necesario decir que el sitio "oficial" desde el que se descargaron muestra sólo once artículos de autoría dudosa, el duodécimo (atribuido a Madison en el primer sitio web) es el artículo número 58 según la *Emory School of Law*[3]. Así, la lista definitiva de artículos disputados sería la siguiente: 49 a 58, 62 y 63.

Artículo	Hamilton	Madison	Autor
49	0,286	0,370	Madison
50	0,282	0,352	Madison
51	0,287	0,370	Madison
52	0,283	0,369	Madison
53	0,298	0,368	Madison
54	0,257	0,375	Madison
55	0,309	0,350	Madison
56	0,255	0,384	Madison
57	0,307	0,367	Madison
58	0,279	0,361	Madison
62	0,291	0,360	Madison
63	0,279	0,368	Madison

Tabla 14. Valores *PiRoNorm* obtenidos por cada texto disputado al compararlo con ambos autores.

Para el conjunto de documentos escritos por cada autor en solitario se calculó el centroide y se empleó éste como vector representativo de la categoría. Como medida de

[1] Programación lineal (Bosch y Smith 1992, citado por Fung 2003), redes neuronales (Tweedie, Singh y Holmes 1994), algoritmos genéticos (Holmes y Forsyth 1994, citado por Buckland 1999), cadenas de Markov (Khmelev y Tweedie 2001) o *SVM* (Fung 2003).

[2] `http://thomas.loc.gov/home/histdox/fedpapers.html`

[3] `http://www.law.emory.edu/FEDERAL/federalist/`

similitud entre los documentos disputados y la categoría se empleó la versión normalizada de *PiRo* (véase pág. 118) y se obtuvieron los resultados que se muestran en Tabla 14 y que permiten atribuir a Madison todos los textos en disputa, lo cual está de acuerdo con las teorías más aceptadas.

6 Filtrado de correo no deseado (*spam*)

El problema del correo no solicitado fue anticipado con gran antelación por Postel (1975) y Peter Denning (1982) señaló la necesidad de investigar no sólo métodos de generar información sino también técnicas que permitan *"controlar y filtrar la información que llega a las personas que deben usarla"*. Cranor y Lamacchia (1998) determinaron que, en 1997, alrededor del 10% del correo recibido en una red corporativa era *spam*. Esta cifra no ha hecho sino aumentar; Whitworth y Whitworth (2004) analizan el trasfondo técnico, legal y social del correo no solicitado y presentan datos alarmantes: aproximadamente el 30% del correo entrante de cada usuario y más del 50% del correo transmitido es *spam*.

La solución del problema no es sencilla; después de todo, al desarrollarse técnicas que filtran mejor este correo basura sus autores envían más aún. Por lo que es muy probable que se sigan utilizando sistemas de categorización automática, bien sea en los servidores de difusión (*relaying*), en los de correo o en las aplicaciones cliente (véase Fig. 86).

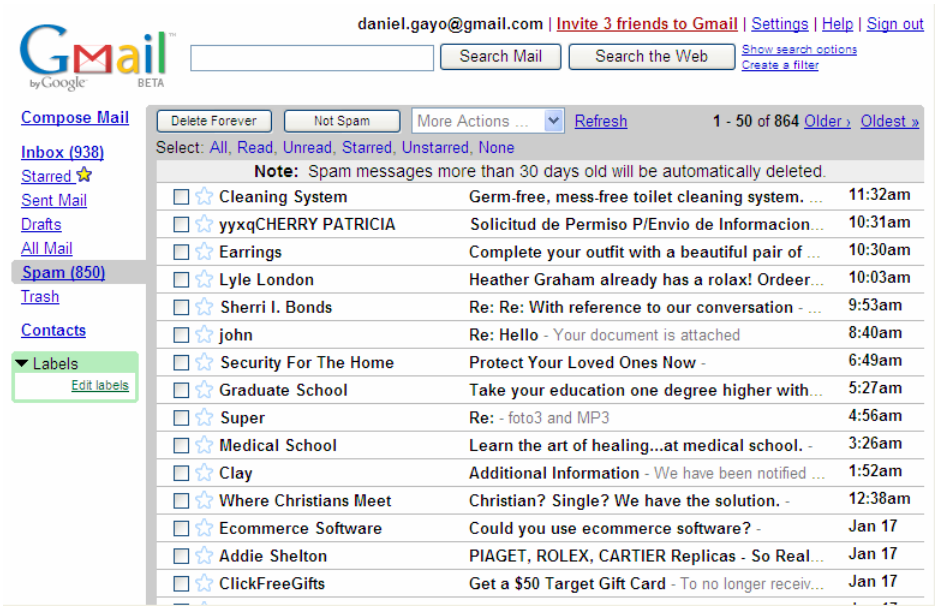


Fig. 86 Correo no solicitado filtrado automáticamente por *Gmail*.

Prácticamente todos los métodos de categorización revisados al comienzo del capítulo se han probado con el *spam*: Sahami *et al.* (1998), Pantel y Lin (1998) y Androutsopoulos *et al.* (2000a) utilizaron categorizadores bayesianos, Drucker, Wu y Vapnik (1999) *SVM's*, Androutsopoulos *et al.* (2000b) usaron *Memory Based Learning* y Zhang y Yao (2003) el modelo de entropía máxima. Según Zhang, Zhu y Yao (2004) *SVM's*, *boosting* y el modelo de máxima entropía son mucho mejores que los categorizadores bayesianos, una de las técnicas mejor consideradas para filtrar *spam* (Androutsopoulos *et al.* 2000a).

A fin de evaluar la utilidad de *blindLight* en esta misión, se emplearon dos *corpora* de correo utilizados en algunos de los trabajos anteriores: `ling-spam`[1] y `spamassassin`[2]. Existen otras colecciones de prueba pero para proteger la privacidad de los usuarios "donantes" no se ofrecen como texto plano sino codificadas haciéndolas inútiles para nuestros propósitos.

La colección `ling-spam` contiene 2.142 mensajes legítimos procedentes de una lista de correo y 481 mensajes no deseados, todos ellos en inglés y sin las correspondientes cabeceras. La colección `spamassassin` consta de 4.150 mensajes legítimos y 1.897 mensajes no deseados, también en inglés e incluyendo cabeceras. La primera colección está dividida en 10 partes para llevar a cabo validación cruzada mientras que la segunda fue dividida por el propio autor.

Para evaluar el rendimiento de la técnica propuesta en la categorización de correo no deseado se emplearán las mismas medidas que Androutsopoulos *et al.* (2000): la exactitud y tasa de error ponderadas, la razón de coste total (*total cost ratio* o *TCR*) así como la precisión y exhaustividad.

Si N_{legit} y N_{spam} es el número de mensajes legítimos y no deseados respectivamente y $n_{Y \to Z}$ es el número de mensajes de la categoría Y asignados a la categoría Z donde Y y Z pueden ser *legit* o *spam* entonces la exactitud (*accuracy*) y la tasa de error serían:

$$Acc = \frac{n_{legit \to legit} + n_{spam \to spam}}{N_{legit} + N_{spam}} \qquad Err = \frac{n_{legit \to spam} + n_{spam \to legit}}{N_{legit} + N_{spam}}$$

Esta definición de error y exactitud otorgan la misma importancia a categorizar un documento legítimo como *spam* que el caso contrario cuando el primer suceso es λ veces más costoso[3] por lo que es necesario ponderar estas medidas (Androutsopoulos *et al.* 2000):

$$WAcc = \frac{\lambda \cdot n_{legit \to legit} + n_{spam \to spam}}{\lambda \cdot N_{legit} + N_{spam}} \qquad WErr = \frac{\lambda \cdot n_{legit \to spam} + n_{spam \to legit}}{\lambda \cdot N_{legit} + N_{spam}}$$

Los valores comunmente asignados a λ son 1, 9 y 999 que se corresponderían a tres escenarios (Androutsopoulos *et al.* 2000): (1) marcar los mensajes no solicitados, (2) enviar una notificación al remitente y (3) borrar los mensajes. Quizás fuese más razonable reemplazar el segundo escenario por uno más plausible como mover los mensajes a una carpeta específica; después de todo, los remitentes de auténtico *spam* no emplean direcciones de correo reales.

No obstante, y a pesar de esta ponderación, la exactitud obtenida suele ser engañosamente alta y se hace necesaria una medida más "intuitiva" que permita comparar el filtro desarrollado con un sistema básico consistente en no filtrar ningún mensaje. La exactitud y tasa de error ponderadas para ese método serían las siguientes.

---

[1] `http://iit.demokritos.gr/skel/i-config/downloads/`

[2] `http://spamassassin.apache.org/publiccorpus/`

[3] Por ejemplo, resulta más sencillo eliminar un correo no deseado no capturado por el filtro que tener que explorar la carpeta de *spam* en busca de algún mensaje legítimo filtrado por error.

$$WAcc^b = \frac{\lambda \cdot N_{legit}}{\lambda \cdot N_{legit} + N_{spam}} \qquad WErr^b = \frac{N_{spam}}{\lambda \cdot N_{legit} + N_{spam}}$$

El cociente entre la tasa de error del sistema básico y del sistema a evaluar es la razón de coste total o *TCR* que tiene una interpretación muy simple: compara el esfuerzo dedicado a eliminar manualmente todo el *spam* recibido frente a eliminar el *spam* que aún pasa el filtro más recuperar correo legítimo categorizado como no solicitado. A mayor valor de *TCR* mejor rendimiento y, por otro lado, un valor de *TCR* inferior a la unidad significa que, en ese escenario en particular, es preferible dejar pasar todo el correo recibido que usar el filtro objeto de análisis.

$$TCR = \frac{WErr^b}{WErr} = \frac{N_{spam}}{\lambda \cdot n_{legit \to spam} + n_{spam \to legit}}$$

Los resultados obtenidos por *blindLight* con la colección `ling-spam` se muestran en Tabla 15. Por lo que respecta a exhaustividad y precisión la técnica propuesta por el autor fue capaz de capturar el 73,59% del *spam* de la colección con una precisión del 96,32% y tan sólo falla en el escenario más exigente (λ=999).

λ	WAcc	TCR
1	95,26%	3,51
9	99,02%	2,22
999	**99,58%**	**0,05**

Tabla 15. Rendimiento de *blindLight* categorizando la colección `ling-spam`.

Los resultados alcanzados al procesar la colección `spamassassin` fueron muy inferiores (véase Tabla 16). La exhaustividad ha sido algo mayor (77,95%) pero la precisión lograda ha sido mucho menor (84,81%). Por lo que respecta al *TCR* este es de 2,78 para λ=1 y de sólo 0,68 para λ=9, esto es, dada esta colección, este escenario y el método implementado sería preferible no filtrar el correo. Es preciso señalar, no obstante, que al procesar `spamassassin` los categorizadores bayesianos requieren vectores de entre 2000 y 3000 términos para conseguir mejorar al método básico (Zhang *et al.* 2004, p.10) no llegando a superar nunca un *TCR* de 2, mientras que con la colección `ling-spam` no precisaban más de 100 características (Androutsopoulos *et al.* 2000).

λ	WAcc	TCR
1	88,70%	2,78
9	**92,86%**	**0,68**
999	**93,91%**	**0,01**

Tabla 16. Rendimiento de *blindLight* categorizando la colección `spamassassin`.

La comparación con otras técnicas ha sido posible, hasta cierto punto, gracias a los trabajos de Androutsopoulos *et al.* (2000) y Zhang *et al.* (2004). Es necesario señalar que la técnica del autor no supera a ninguno de los otros métodos; sin embargo, en el caso de los categorizadores bayesianos y los que emplean *memory based learning* tan sólo hay diferencias en la exhaustividad, es decir, dichos métodos capturan más *spam* que *blindLight*. Estas diferencias son materiales en el caso de λ=1 y sólo apreciables para λ=9 y únicamente con el categorizador bayesiano. Métodos como *SVMs*, *boosting* o el de máxima entropía son muy superiores no sólo a *blindLight* sino también a los categorizadores bayesianos y *MBL* (Zhang *et al.* 2004) aunque sólo se ofrecen datos para λ=9 y 999, no para λ=1.

Se pueden extraer dos conclusiones de estos experimentos. En primer lugar antes de aplicar *blindLight* de forma efectiva a la particular tarea de filtrar *spam* sería necesario determinar la forma de introducir un modo de ponderar de manera distinta ambas categorías. Por otro lado, analizando estos datos como un experimento de categorización más, esto es, centrándonos en el escenario de $\lambda=1$, tenemos que la utilización de *blindLight* como categorizador proporciona resultados próximos a los de los del método bayesiano y *MBL*.

7 Comparación de *blindLight* con otras técnicas de categorización

Para poder comparar el rendimiento de la técnica propuesta por el autor con otros métodos de categorización existentes se llevó a cabo una serie de experimentos sobre las colecciones `Reuters-21578`[1] y `OHSUMED`[2].

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="5549"
NEWID="6">
  <DATE>26-FEB-1987 15:14:36.41</DATE>
  <TOPICS>
    <D>veg-oil</D><D>linseed</D><D>lin-oil</D><D>soy-oil</D><D>sun-oil</D>
    <D>soybean</D><D>oilseed</D><D>corn</D><D>sunseed</D><D>grain</D>
    <D>sorghum</D><D>wheat</D>
  </TOPICS>
  <PLACES>
    <D>argentina</D>
  </PLACES>
  <PEOPLE></PEOPLE>
  <ORGS></ORGS>
  <EXCHANGES></EXCHANGES>
  <COMPANIES></COMPANIES>
  <UNKNOWN>
    G   f0754 reuter f BC-ARGENTINE-1986/87-GRA   02-26 0066
  </UNKNOWN>
  <TEXT>
    <TITLE>
      ARGENTINE 1986/87 GRAIN/OILSEED REGISTRATIONS
    </TITLE>
    <DATELINE>
      BUENOS AIRES, Feb 26 -
    </DATELINE>
    <BODY>
      Argentine grain board figures show crop registrations of grains,
      oilseeds and their products to February 11, in thousands of tonnes,
      showing those for futurE shipments month, 1986/87 total and 1985/86
      total to February 12, 1986, in brackets:
      Bread wheat prev 1,655.8, Feb 872.0, March 164.6, total
      2,692.4 (4,161.0).
      Maize Mar 48.0, total 48.0 (nil).
      Sorghum nil (nil)
      Oilseed export registrations were:
      Sunflowerseed total 15.0 (7.9)
      Soybean May 20.0, total 20.0 (nil)
    </BODY>
  </TEXT>
</REUTERS>
```

Fig. 87 Un documento de la colección *Reuters-21578*.

La primera consta de una serie de artículos (véase Fig. 87) publicados por la agencia de prensa *Reuters* durante el año 1987 a los que se asignaron manualmente una o más "etiquetas" de una lista[3] de 135 posibles. En la literatura se han empleado distintas particiones de la colección en conjuntos de entrenamiento y prueba. Por tanto, para obtener resultados comparables con los obtenidos por Joachims (1997) y Dumais *et al.* (1998) se ha

[1] `http://www.daviddlewis.com/resources/testcollections/reuters21578/`

[2] `ftp://medir.ohsu.edu/pub/ohsumed`

[3] En realidad, la colección `Reuters-21578` proporciona 5 conjuntos de categorías (`Exchanges`, `Orgs`, `People`, `Places` y `Topics`). Sin embargo, se suelen emplear únicamente las categorías correspondientes a `Topics` que hacen referencia a asuntos económicos, por ejemplo: `crude`, `nat-gas` o `iron-steel`.

utilizado la misma partición que estos investigadores, la denominada ModApte que utiliza 9.603 documentos para entrenamiento, 3.299 para test y descarta el resto de la colección.

Bacterial Infections and Mycoses	C01
Virus Diseases	C02
Parasitic Diseases	C03
Neoplasms	C04
Musculoskeletal Diseases	C05
Digestive System Diseases	C06
Stomatognathic Diseases	C07
Respiratory Tract Diseases	C08
Otorhinolaryngologic Diseases	C09
Nervous System Diseases	C10
Eye Diseases	C11
Urologic and Male Genital Diseases	C12
Female Genital Diseases and Pregnancy Complications	C13
Cardiovascular Diseases	C14
Hemic and Lymphatic Diseases	C15
Neonatal Diseases and Abnormalities	C16
Skin and Connective Tissue Diseases	C17
Nutritional and Metabolic Diseases	C18
Endocrine Diseases	C19
Immunologic Diseases	C20
Disorders of Environmental Origin	C21
Animal Diseases	C22
Pathological Conditions, Signs and Symptoms	C23

Fig. 88 Las 27 categorías de enfermedades presentes en *MeSH*.

Por lo que respecta a la colección OHSUMED consta de 348.566 referencias extraídas de *MEDLINE* y compuestas por el título y/o el resumen de artículos publicados en revistas médicas entre 1987 y 1991. Cada documento de la colección tiene asignado uno más términos de la clasificación *MeSH* (*Medical Subject Headings*) que proporciona un vocabulario controlado que abarca aspectos tan diversos como anatomía, trastornos mentales, procedimientos o medicamentos. Joachims (1997) utilizó los 20000 primeros documentos correspondientes a 1991 que incluían resumen, empleando la primera mitad durante el entrenamiento y la segunda para la fase de prueba. En cuanto a las categorías se limitó a las 23 entradas de primer nivel correspondientes a enfermedades (véase Fig. 88) suponiendo que un documento pertenece a una categoría dada si tiene asociado al menos un término índice de dicha categoría (véase Fig. 89).

```
Abdominal Pain/ET; Adolescence; Adult; Aged; Aged, 80 and over;
Appendicitis/CO/*RI/US; Child; Female; Human; Leukocytes/*; Middle Age;
Predictive Value of Tests; Support, Non-U.S. Gov't; Technetium Tc 99m
Aggregated Albumin/*DU.
```

```
Abdominal Pain;C23.888.646.100
Abdominal Pain;C23.888.821.030
Adult;M01.060.116
Aged;M01.060.116.100
Aged, 80 and over;M01.060.116.100.080
Appendicitis;C06.405.205.099
Appendicitis;C06.405.469.110.207
Child;M01.060.406
Leukocytes;A11.118.637
Leukocytes;A15.145.229.637
Predictive Value of Tests;E05.318.780.800.650
Predictive Value of Tests;G03.850.520.445.800.650
Predictive Value of Tests;H01.548.832.672.500
Predictive Value of Tests;N05.715.360.780.700.640
Technetium Tc 99m Aggregated Albumin;D02.691.825.375
Technetium Tc 99m Aggregated Albumin;D12.776.034.900
```

Fig. 89 Términos asociados a un documento de la colección *OHSUMED* y categorías *MeSH* correspondientes.

Joachims (1997), Dumais *et al.* (1998) y otros investigadores citados por Sebastiani (2002) emplearon como indicador del rendimiento de las técnicas analizadas el *breakeven*[1], aquel punto en el cual precisión y exhaustividad son iguales; la definición de precisión y exhaustividad en este contexto se presenta en las siguientes ecuaciones:

$$exhaustividad = \frac{categorías\ correctamente\ asignadas}{total\ de\ categorías\ correctas}$$

$$precisión = \frac{categorías\ correctamente\ asignadas}{total\ de\ categorías\ asignadas}$$

Así, al llevar a cabo los experimentos con *blindLight* se obtuvo, para cada documento, una lista ordenada de categorías comenzando por las más fuertemente vinculadas y terminando con las menos relacionadas con el documento. A partir de estos datos se extrajeron los correspondientes a precisión y exhaustividad que fueron "interpolados" (Chakrabarti 2003, p. 55) y, posteriormente, micropromediados. Una vez obtenidos estos últimos se aplicó interpolación lineal a aquellos pares de valores que delimitaban el valor de *breakeven* buscado.

Los resultados obtenidos por la técnica del autor en ambas colecciones se muestran en Tabla 17 y Tabla 18 junto con los alcanzados por Joachims (1997) y Dumais *et al.* (1998). La Tabla 19 permite comparar *blindLight* con toda una serie de técnicas aplicadas para categorizar la partición `ModApte` de la colección `Reuters-21578`. A la vista de tales resultados puede concluirse que, en general, *blindLight* es capaz de alcanzar resultados análogos a los de Rocchio, categorizadores bayesianos o árboles de decisión, resultados próximos aunque apreciablemente inferiores a los de *k*-vecinos y sustancialmente inferiores a los obtenidos con *SVMs*.

	blindLight	Bayes (i)	Bayes (ii)	Redes Bayes	Rocchio	*Findsim*	C4.5	Árboles de decisión	k-NN	SVM (poly) *d*=4	SVM (RBF) γ=0,8	SVM (lineal)
earn	94,5%	95,9%	95,9%	95,8%	96,1%	92,9%	96,1%	97,8%	97,3%	98,4%	**98,5%**	98,0%
acq	**99,3%**	91,5%	87,8%	88,3%	92,1%	64,7%	85,3%	89,7%	92,0%	95,2%	95,3%	93,6%
money-fx	51,7%	62,9%	56,6%	58,8%	67,6%	46,7%	69,4%	66,2%	**78,2%**	74,9%	75,4%	74,5%
grain	62,9%	72,5%	78,8%	81,4%	79,5%	67,5%	89,1%	85,0%	82,2%	91,3%	91,9%	**94,6%**
crude	76,3%	81,0%	79,5%	79,6%	81,5%	70,1%	75,5%	85,0%	85,7%	88,9%	**89,0%**	88,9%
trade	**95,5%**	50,0%	63,9%	69,0%	77,4%	65,1%	59,2%	72,5%	77,4%	77,3%	78,0%	75,9%
interest	39,6%	58,0%	64,9%	71,3%	72,5%	63,4%	49,1%	67,1%	74,0%	73,1%	75,0%	**77,7%**
ship	52,2%	78,7%	85,4%	84,4%	83,1%	49,2%	80,9%	74,2%	79,2%	86,5%	**86,5%**	85,6%
wheat	38,1%	60,6%	69,7%	82,7%	79,4%	68,9%	85,5%	**92,5%**	76,6%	85,9%	85,9%	91,8%
corn	31,3%	47,3%	65,3%	76,4%	62,2%	48,2%	87,7%	**91,8%**	77,9%	85,7%	85,7%	90,3%
μpromedio	77,7%	72,0%	75,2%	80,0%	79,9%	61,7%	79,4%	¿?	82,3%	86,2%	86,5%	87,0%

Tabla 17. Comparación de *blindLight* con otras técnicas al categorizar la colección *Reuters-21578*.

Se muestran los resultados para las 10 categorías más frecuentes y micropromediados para todas las categorías. Los datos de *Bayes* (i), *Rocchio*, *C4.5*, *k-NN*, *SVM* (*poly* y *RBF*) pertenecen a Joachims (1997), el resto a Dumais *et al.* (1998).

	blindLight	Bayes	Rocchio	C4.5	k-NN	SVM (poly) *d*=4	SVM (RBF) γ=1,0
Pathological Conditions, Signs and Symptoms	**83,9%**	52,7%	50,8%	47,6%	53,4%	58,2%	58,1%
Cardiovascular Diseases	69,6%	72,4%	70,1%	70,5%	72,6%	77,3%	**77,6%**
Immune System Diseases	55,0%	61,7%	58,0%	58,8%	66,8%	73,2%	**73,5%**
Neoplasms	**77,0%**	63,6%	64,1%	58,7%	67,2%	70,6%	70,7%
Digestive System Diseases	33,5%	65,3%	59,9%	59,0%	67,1%	73,7%	**73,8%**
Micropromedio	54,0%	57,0%	56,6%	50,0%	59,1%	65,9%	66,1%

Tabla 18. Comparación de *blindLight* con otras técnicas al categorizar la colección *OHSUMED*.

Se muestran los resultados para las 5 categorías más frecuentes y micropromediados para todas las categorías. Todos los datos proceden de Joachims (1997).

[1] Tanto Joachims (1997) como Dumais *et al.* (1998) señalan que los datos fueron micropromediados.

Técnica	Rendimiento
Boosting	87,8%
SVM	85,9%
k-NN	84,0%
Redes neuronales	83,8%
Reglas decision	82,2%
Redes Bayes	80,0%
C4.5	79,4%
blindLight	**77,7%**
Bayes	75,7%
Rocchio	72,0%

Tabla 19. Comparación del rendimiento de *blindLight* con otras técnicas al categorizar la partición *ModApte* de la colección *Reuters-21578*. Los datos se han obtenido de Sebastiani (2002, p. 38).

8 Influencia del tamaño de los *n*-gramas en la categorización

Todos los experimentos descritos hasta el momento fueron llevados a cabo empleando vectores de 3-gramas. En el capítulo anterior se llevó a cabo un experimento para evaluar la influencia del tamaño de los *n*-gramas sobre los resultados de la clasificación automática y en éste se ha hecho lo propio para determinar la influencia sobre el rendimiento del categorizador. Para ello se repitió el experimento descrito para la colección OHSUMED con 2-, 3- y 4-gramas.

Los resultados obtenidos se muestran en Tabla 20 y de ellos se desprende que el rendimiento al emplear vectores de bigramas en tareas de categorización no es adecuado al compararlo con los resultados obtenidos con vectores de 3- o 4-gramas. Sin embargo, para estos dos últimos tamaños no existen diferencias apreciables por lo que sería recomendable emplear trigramas en tareas de categorización puesto que tanto el tiempo de procesamiento como el espacio de almacenamiento necesarios son mucho menores.

Baste decir para terminar que se ha mostrado cómo la técnica propuesta por el autor puede efectivamente aplicarse al problema de la categorización de texto libre y obtener resultados análogos a los de muchos de los métodos convencionales.

	2-gramas	3-gramas	4-gramas
Pathological Conditions, Signs and Symptoms	44,03%	83,89%	**86,59%**
Cardiovascular Diseases	32,01%	**69,64%**	66,89%
Immune System Diseases	27,22%	**54,98%**	53,36%
Neoplasms	**86,95%**	76,95%	76,40%
Digestive System Diseases	**39,73%**	33,54%	37,39%
Micropromedio	51,45%	53,96%	**54,42%**
Macropromedio (top 5)	45,99%	63,80%	**64,13%**

Tabla 20. Comparación del rendimiento de *blindLight* empleando distintos tamaños de *n*-grama.

CAPÍTULO 6

RECUPERACIÓN DE INFORMACIÓN CON *BLINDLIGHT*

El término recuperación de información (IR) hace referencia, en general, al estudio de sistemas automáticos que permitan a un usuario determinar la existencia o inexistencia de documentos (esto es, textos) relativos a una necesidad de información formulada habitualmente como un pequeño fragmento de texto conocido como "consulta" (una o más frases o una simple secuencia de "palabras clave"). Los orígenes de este campo pueden remontarse a los años 1950 y desde entonces ha madurado enormemente, en particular en lo tocante a la evaluación sistemática de los sistemas IR. En la actualidad la recuperación de información está caracterizada por tres atributos dispares: su naturaleza interactiva, el número de documentos a tratar y el carácter multilingüe de documentos y usuarios. En este capítulo se analizará el uso de blindLight como técnica de recuperación de información, las características de la misma que la hacen especialmente interesante en entornos multilingües y los resultados obtenidos en pruebas estandarizadas. Así mismo, se compararán estos resultados con los alcanzados por otras técnicas IR y se señalarán distintas líneas de trabajo futuro que se espera contribuyan a mejorar esta nueva técnica con el objeto de ponerla al mismo nivel que otros métodos de recuperación de información ya afianzados.

1 Recuperación de información

El término "recuperación de información" (*information retrieval* o *IR*) hace referencia al conjunto de procesos necesarios para representar, almacenar, buscar y encontrar información relevante para las consultas de los usuarios (Ingwersen 1992, p. 49). A pesar de su carácter central en el ámbito de las ciencias y tecnologías de la información (Griffith 1980, p. 239) (Järvelin y Vakkari 1992, citado por Ingwersen 1992) se trata de un término vagamente definido (van Rijsbergen 1979, p. 1) puesto que, en principio, podría referirse a diversas manifestaciones de la información como imágenes, audio, texto, etc. No obstante, se acepta generalmente que la "recuperación de información" se ocupa únicamente de información textual (Ingwersen 1992, p. 50) y describe sistemas análogos a los desarrollados,

por ejemplo, por Salton, Spärck-Jones o Robertson. Es interesante en este sentido la definición propuesta por Lancaster (1968, citado por Rijsbergen 1979, p. 1):

> *Recuperación de información es el término que se aplica habitualmente, aunque de manera inexacta, al tipo de trabajo descrito en esta obra*[1]*. Un sistema de recuperación de información no informa al usuario acerca del tema de su consulta, es decir, no modifica sus conocimientos. Simplemente, indica la existencia (o inexistencia) y localización de documentos relativos a dicha consulta.*

Se trata de una definición conveniente puesto que describe a la perfección toda una serie de sistemas desarrollados durante un período de tiempo muy amplio: desde *SMART* (Salton y Lesk 1965) (Salton 1968) hasta *Google* (Brin y Page 1998) por fijar sólo dos hitos. Sin embargo, no tiene en cuenta un aspecto muy importante de la *IR*: su naturaleza interactiva. En fecha muy temprana Don Swanson[2] (1977) señaló:

> *…la recuperación de información es un proceso de ensayo y error… Una consulta no es más que una suposición acerca de los atributos que se espera tenga el documento deseado. En general, se emplea la respuesta del sistema para corregir esa suposición inicial en posteriores intentos.*

No obstante, en los primeros momentos de la investigación en sistemas *IR* no se establece la interactividad como un requisito sino que se trata de una característica que emerge con la evolución[3] de estos sistemas. Los aspectos interactivos e iterativos en los procesos de recuperación de información no comenzaron a estudiarse hasta los años 80 –por ejemplo, Belkin y Vickery (1985), Croft y Thompson (1987), Bates (1989) o Ingwersen (1992)– y fueron totalmente aceptados en los años 90 reconociéndose la necesidad de investigar no sólo los resultados de los sistemas *IR* sino la forma en que son utilizados por los usuarios (Harman 1996).

Así pues es posible distinguir tres enfoques (Ingwersen 1992) en la investigación de sistemas *IR*: el "tradicional", el "orientado al usuario" y el "cognitivo". En el primer caso el objeto de estudio es la representación de documentos y consultas así como las funciones de "emparejamiento" entre ambos tipos de textos. El enfoque "orientado al usuario" se centra en los componentes humanos de un sistema interactivo de recuperación de información. Por último, el enfoque "cognitivo" trata de desarrollar un enfoque integrador de todos los componentes (automatizados y humanos) de un sistema completo.

En este trabajo se ha estudiado la aplicación de *blindLight* a la recuperación de información desde un enfoque "tradicional". Ciertamente, muchas de las aplicaciones de esta técnica (clasificación, categorización o extracción de resúmenes) podrían ser muy útiles en un sistema interactivo y sería muy interesante estudiar su aplicación. Sin embargo, el autor consideró que este análisis no era uno de los puntos a tratar en esta disertación.

A lo largo de los siguientes apartados se repasará brevemente la evolución de los sistemas *IR*, se estudiará la forma en que se realiza su evaluación (desde el enfoque

---

[1] Lancaster, F.W. 1968, *Information Retrieval Systems: Characteristics, Testing and Evaluation.*

[2] No obstante, quizás la contribución fundamental de Swanson haya sido la propuesta de técnicas para generar hipótesis médicas de manera semi-automática a partir de colecciones bibliográficas (Swanson 1986) (Swanson 1991) (Swanson y Smalheiser 1997) (Smalheiser y Swanson 1998). La última implementación de tales técnicas está disponible en `http://arrowsmith.psych.uic.edu/arrowsmith_uic`.

[3] Según Salton y Crouch (1989, p. 3) son las mejoras en el *hardware* y las interfaces gráficas de usuario las que han posibilitado *"sofisticadas interacciones entre los usuarios y los sistemas [de recuperación de información]."*

tradicional), se describirá la forma en que es posible aplicar *blindLight* a esta tarea y se analizarán los resultados obtenidos con distintas colecciones y en la participación en *CLEF'04*[1].

2 Evolución de los sistemas de recuperación de información

Según Herbert Ohlman (1999) las tecnologías de procesamiento del lenguaje tal y como las entendemos en la actualidad comenzaron su andadura a mediados del S. XX. En esta época se desarrollaron técnicas para construir concordancias[2] (véase Fig. 90) como el "indexado por permutación" (*permutation indexing*) (Ohlman 1957) o la técnica análoga *KWIC* (*keyword in context*) de Luhn (1959, citado por Ohlman 1999), se propusieron técnicas para obtener resúmenes automáticos (Luhn 1958) y, por supuesto, se dieron los primeros pasos hacia los actuales sistemas de recuperación de información (Luhn 1957) y (Baxendale 1958).

cenamiento (600 Mb), la rapidez de **recuperación de información**, la posibilidad de realiz**
es implicaciones en el campo de la **recuperación de información** y de la valoración sobre **
rld Wide Web, sistema basado en la **recuperación de información** a partir de técnicas de h**
. Servicio Internet Localización y **recuperación de información** relacionada con las Cienc**
ctualidad, como la extracción y la **recuperación de información**. Además de comunicaciones**
üística de corpus. 3. Extracción y **recuperación de información**. 4. Gramáticas y formalis**
tauración especiales tales como la **recuperación de información** de seguridad, restauració**
as). - Desarrollar los sistemas de **recuperación de información** clínica para usos asisten**

Fig. 90 Concordancias extraídas del *Corpus de Referencia del Español Actual* de la *RAE*.

Luhn (1957, p. 313) planteó lo que podría considerarse el núcleo básico de los sistemas de recuperación de información:

> *Cuanto mayor sea la coincidencia entre los elementos de dos representaciones y entre las distribuciones de los mismos, mayor será la probabilidad de que representen información similar.*

Para implementar esta idea en un sistema viable de recuperación de información Luhn propone, en primer lugar, extraer concordancias para las distintas palabras de una colección a fin de utilizar esos datos para construir (mediante expertos) familias de "nociones". Posteriormente, dichas nociones serían utilizadas para codificar los documentos de la colección, aunque sólo las nociones principales (las más frecuentes o las utilizadas en títulos, encabezados y resúmenes) aparecerían en la representación final de cada documento.

Para llevar a cabo las consultas Luhn sugiere que el usuario proporcione un documento en el que describa de la forma más detallada posible la naturaleza del problema para el que pretende hallar respuesta. Este documento sería codificado de la misma manera que los pertenecientes a la colección y comparado con éstos: a mayor número de nociones en común mayor similitud entre el documento y la consulta. Luhn también apunta la posibilidad de "expandir" automáticamente la consulta original mediante nociones relacionadas.

Luhn no proporcionó ningun resultado empírico, tan sólo afirmó que un experimento llevado a cabo con 1200 informes técnicos produjo "resultados esperanzadores" y, a la vista de los últimos 50 años, sin duda lo eran. Por otro lado, es importante señalar que Luhn introdujo algunos conceptos relevantes en este campo como la

[1] Recuérdese que el *CLEF – Cross Language Evaluation Forum* es un foro internacional para la evaluación de sistemas de recuperación de información que operen sobre idiomas europeos.

[2] Índice de todas las palabras de un libro o del conjunto de la obra de un autor, con todas las citas de los lugares en que se hallan (*RAE* 2001).

utilización de la frecuencia de los términos, del marcado de partes del habla[1] o la expansión de consultas mediante diccionarios de sinónimos.

El trabajo de Baxendale (1958) se centró en la extracción automática de términos clave para su utilización como índices. Baxendale propuso tres métodos para realizar dicha tarea señalando una serie de puntos importantes: la existencia de "palabras vacías", la relación entre "significatividad" y frecuencia de aparición y la dificultad de evaluar la calidad de los términos extraídos automáticamente al compararlos con otros propuestos por expertos.

Los trabajos anteriores plantearon ideas extremadamente interesantes pero fueron Maron y Kuhns (1960) los primeros en proponer un sistema de recuperación que incorporase de una manera efectiva el concepto de "relevancia" de un documento.

Luhn (1958, p. 313) ya había señalado que el grado de coincidencia entre los términos y entre las distribuciones de los mismos en dos documentos sería un indicador de la probabilidad de que ambos documentos traten temas similares. Sin embargo, no describió ninguna técnica para calcular dichas probabilidades automáticamente.

Por otro lado, Baxendale (1958) estaba interesado en reducir cada documento a un número de términos índice pequeños y, aunque no detalló el modo en que se haría la recuperación, cabe aventurar que la técnica empleada sería una búsqueda booleana. Al emplearse muy pocos términos índice por documento los resultados serían reducidos; sin embargo, se plantearían problemas si los usuarios empleasen términos similares aunque distintos de los extraídos del texto de los documentos[2].

Maron y Kuhns señalaron dos aspectos importantes en el campo de recuperación de información: (1) la idea de relevancia como una cantidad numérica que aunque puede carecer de valor como medida cuantitativa sí resulta útil en términos comparativos y (2) el hecho de que una consulta es, por naturaleza, imprecisa, tan sólo una "pista" sobre las necesidades de información del usuario y que el propio sistema *IR* debe "elaborar" dicha pista.

Maron y Kuhns argumentan que conocida la probabilidad de que un término sea aplicado a un documento y sabida la frecuencia de acceso a cada documento es posible, aplicando el teorema de Bayes, calcular la probabilidad de que un documento en particular sea considerado relevante para un término dado y por extensión para una consulta que combina varios términos. Por otro lado, describen una serie de técnicas que permitirían expandir consultas añadiendo nuevos términos relacionados con los empleados por el usuario, así como la forma de ponderarlos. Aproximadamente en la misma época Lauren B. Doyle (1959 y 1965, citado por van Rijsbergen 1979, p. 108) y H.E. Stiles (1961, citado por van Rijsbergen 1979, p. 108) utilizaron de manera similar la co-ocurrencia de términos.

Además de esto, Maron y Kuhns fueron los primeros en proporcionar resultados empíricos y demostrar de manera rigurosa que un sistema de recuperación de información automatizado era factible. Es cierto que no se trataba de búsquedas en "textos completos" y que los términos índice eran extraídos y ponderados manualmente pero la importancia de las ideas planteadas en su trabajo es indudable.

---

[1] *Sería ventajoso identificar mediante símbolos especiales ciertas clases [de palabras] como sustantivos, adjetivos o nombres.* (Luhn 1957, p. 314)

[2] Posteriormente, Lewis, Baxendale y Bennett (1967) investigarían la forma de determinar estadísticamente relaciones de sinonimia/antonimia.

El modelo vectorial de documentos fue utilizado por primera vez en el sistema *SMART* (Salton y Lesk 1965) y ya se ha descrito con detalle en la página 45 y posteriores, baste tan sólo decir que en este modelo los documentos son representados como vectores en un espacio *n*-dimensional donde los términos son empleados como coordenadas a las que se asigna un peso calculado a partir de la frecuencia de uso de cada término en el propio documento y en la colección completa. Para llevar a cabo la recuperación de documentos relevantes para una consulta es necesario representar dicha consulta como un vector y calcular la similitud entre el vector consulta y los vectores documento empleando medidas de asociación como la función del coseno.

Es necesario decir que el modelo vectorial es más bien una familia de técnicas de recuperación de información: por un lado pueden emplearse toda clase de elementos como términos (*n*-gramas, palabras, raíces, lemas, etc.) y, por otro, aplicarse distintos esquemas de ponderación para dichos términos. Así, una de las variantes consideradas más efectivas es la que emplea *tf*idf* junto con la denominada *pivoted document length normalization*[1] (Singhal, Buckley y Mitra 1996) que trata de evitar la tendencia de la medida coseno a favorecer la recuperación de los documentos más cortos de la colección. Por otro lado, el modelo vectorial también admite la expansión automática de consultas sugerida por Maron y Kuhns (1960): Rocchio (1966) introdujo una técnica que permitía ampliar una consulta original empleando para ello la información proporcionada por el usuario al señalar los documentos relevantes dentro del conjunto de resultados[2].

Otro modelo para recuperación de información es el probabilista del que Maron y Kuhns (1960) fueron pioneros. No obstante, serían Karen Spärck-Jones y Stephen Robertson (1976) los que sentarían unas bases realmente sólidas para su utilización efectiva. En la propuesta de Maron y Kuhns era vital conocer la probabilidad con que un término sería utilizado como índice de un documento dado; sin embargo, dicho peso debía ser establecido por un experto humano y el número de términos índice era, por tanto, reducido.

Spärck-Jones (1972) demostró que la especificidad de un término era inversamente proporcional a su frecuencia de uso en la colección, es decir, cuanto mayor es el número de documentos que incluyen un término menos específico resulta como índice y viceversa. Así pues, resultaba posible ponderar cualquier término empleado en una colección y emplearlo como índice de una manera totalmente automática, para un término *t* que apareciese en *n* documentos de una colección formada por *N* documentos el peso sería:

$$w = \log \frac{N}{n}$$

Esta idea sería aplicada, como ya se ha dicho con anterioridad, al método de ponderación conocido como *tf*idf* muy empleado en el modelo vectorial pero, además, daría lugar al mencionado modelo probabilístico que, además de la frecuencia *idf*, utiliza información sobre la relevancia de los documentos para los distintos términos a modo de "entrenamiento" (Robertson y Spärck-Jones 1976) (Spärck-Jones 1979).

---

[1] Normalización de la longitud del documento mediante pivote.

[2] El principal inconveniente de esta técnica es la necesidad de una realimentación explícita que los usuarios son reacios a proporcionar (Balabanovic 1998, p. 6). Una solución a este problema es el denominado *pseudo-relevance feedback* (pseudo-realimentación de relevancia) consistente, *grosso modo*, en la expansión de la consulta original mediante términos extraídos de los primeros documentos obtenidos como resultados. Buckley *et al.* 1994, Robertson *et al.* 1994 o Mitra, Singhal y Buckley 1998 son algunos de los que han mostrado la utilidad de este método.

Este modelo emplea los siguientes parámetros para estimar el peso de un término *t* que aparezca en una consulta *q*: *n* es el número de documentos que incluyen *t*, *N* el número de documentos de la colección, *r* el número de documentos relevantes que incluyen *t* y *R* el número de documentos relevantes para *q*. Robertson y Spärck-Jones proponen distintos esquemas de ponderación aunque el preferido, en particular para situaciones predictivas[1], sería el siguiente:

$$w = \log \frac{(r+0{,}5)(N-n-R+r+0{,}5)}{(R-r+0{,}5)(n-r+0{,}5)}$$

Posteriormente, este modelo sería extendido hasta convertirse en el conocido como *BM25* (Robertson *et al.* 1994) y que es considerado como uno de los métodos *IR* probabilistas más efectivos. Por otro lado es necesario señalar que van Rijsbergen (1977), Harper y van Rijsbergen (1978) o Bookstein y Kraft (1977, citado por van Rijsbergen 1979, p. 108) trabajaron en modelos probabilistas de recuperación de información que no suponen, como el anterior, la independencia entre los términos o que Croft y Harper (1979) desarrollaron un modelo similar pero que no requiere información (explícita o implícita) sobre la relevancia de los documentos.

Naturalmente hay muchos otros modelos para llevar a cabo recuperación de información. Ya se citó el booleano, además del booleano extendido (Salton, Fox y Wu 1983), el vectorial generalizado (Wong, Ziarko y Wong 1985), el basado en conjuntos difusos (Kraft y Buell 1983) o (Cross 1994), o el de semántica latente (Deerwester *et al.* 1990) por citar unos pocos. Para una revisión más profunda de los distintos modelos de *IR* el lector puede acudir al segundo capítulo del excelente *"Modern Information Retrieval"* (Baeza-Yates y Ribeiro-Neto 1999).

3 Evaluación de sistemas de recuperación de información

Un sistema de recuperación de información ideal debería proporcionar tan sólo documentos relevantes para las consultas que recibiese. Sin embargo, en la práctica se acepta que el objetivo de un sistema *IR* es localizar el mayor número posible de documentos relevantes junto con el menor número posible de documentos irrelevantes. Además, el sistema ofrece los resultados de manera ordenada, esto es, cuanto más relevante se presume un documento antes aparecerá en la lista de resultados y viceversa.

Ciertamente, la relevancia de un documento es una cualidad subjetiva que cambia con cada usuario. No obstante, esto no plantea mayores problemas en un marco experimental pues es posible proporcionar una colección de documentos junto con un conjunto de consultas de prueba para las cuales un grupo de expertos decide qué documentos son relevantes. En general, se acepta que si un sistema de recuperación de información funciona de manera adecuada bajo diversas condiciones experimentales también lo hará en condiciones no controladas (van Rijsbergen 1979, p. 113).

Se plantea entonces una serie de cuestiones: la elaboración de colecciones y documentos, las medidas que se tomarán como indicadores del rendimiento, así como el desarrollo del propio experimento. No parece adecuado entrar en excesivos detalles sobre la evaluación en *IR* puesto que el capítulo séptimo de *"Information Retrieval"* (van Rijsbergen 1979) y el resto de referencias de este apartado cubren los distintos aspectos de la evaluación

[1] Obteniendo información sobre la relevancia de los resultados del propio usuario o empleando métodos de pseudo-realimentación de relevancia (que suponen que los primeros *T* resultados son relevantes).

en recuperación de información. Por ello, nos limitaremos a describir someramente las medidas de rendimiento más habituales en *IR* y algunos de los principales esfuerzos hechos para lograr entornos de evaluación válidos.

3.1 ¿Cómo medir el rendimiento de un sistema *IR*?

Dejando a un lado los aspectos interactivos, son dos los parámetros generalmente empleados para evaluar la efectividad de un sistema de recuperación de información: la **exhaustividad** (*recall*) y la **precisión**. La exhaustividad es la proporción de documentos relevantes en la colección que se retornan como respuesta a una consulta mientras que la precisión es la proporción de documentos retornados que son realmente relevantes. Otra medida alternativa es el ***fallout***, la proporción de documentos no relevantes en la colección que aparecen en los resultados (véase Tabla 21).

Estos valores se determinan para cada consulta. Dada una consulta, un sistema *IR* puede retornar una lista de documentos $(d_1, d_2, \ldots, d_k)$ ordenados por relevancia, donde k variaría entre 1 y N, siendo éste el número de documentos en la colección. De este modo se puede determinar la precisión en k y, a partir de dichos valores obtener la denominada **precisión media**. Por otro lado, es posible obtener la llamada **precisión interpolada** para cada consulta en una serie de valores de exhaustividad prefijados (0, 0.1, 0.2, …, 1) y, posteriormente, macropromediar los resultados que se mostrarán en una única **curva precisión-exhaustividad**.

	Recuperados	No recuperados	
Relevantes	w	x	$n_1=w+x$
No relevantes	y	z	$n_2=y+z$
	$n_3=w+y$		$N=w+x+y+z$

Tabla 21. Tabla de "contingencia" en recuperación de información.

En esta tabla *w* es el número de documentos relevantes obtenidos, *x* el número de documentos relevantes que no aparecen en los resultados, *y* el número de documentos irrelevantes que sí aparecen en los resultados y *z* los que no aparecen en los resultados. De este modo la precisión sería w/n_3, la exhaustividad w/n_1 y el fallout y/n_2.

Otra medida que también trata de plasmar la efectividad de un sistema de recuperación de información mediante un valor único es la denominada ***medida F*** de van Rijsbergen (1979, pp. 129-135). En realidad van Rijsbergen propuso un marco en el que podían obtenerse distintas medidas cambiando un parámetro α. La medida original de la efectividad, E, era la siguiente (P es la precisión y R la exhaustividad, *recall*):

$$E = 1 - \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)}$$

La medida F es igual a 1-E de tal modo que valores pequeños están asociados a un rendimiento pobre y valores elevados a un rendimiento alto:

$$F = \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)}$$

De este modo, si α vale 0 la medida *F* equivaldría a la exhaustividad mientras que si valiese 1 pasaría a ser la precisión. No obstante, el valor habitualmente empleado al referirse a esta medida es α=½ con lo que la medida *F* queda como:

$$F = 2\frac{P \cdot R}{P + R}$$

3.2 Hitos en la evaluación de los sistemas *IR*

Según Harman (1993) el origen de la evaluación experimental en *IR* puede remontarse al trabajo de Cleverdon (1962, citado por Harman 1993) en el proyecto *Cranfield I* donde se evaluaron distintos lenguajes de indexado. Posteriormente, Cleverdon *et al.* (1966) demostraron que las técnicas de indexado automático proporcionaban resultados análogos al indexado manual y sentarían las bases de la evaluación en recuperación de información: la creación de colecciones de documentos, conjuntos de consultas y subconjuntos de documentos relevantes a fin de facilitar la comparación entre distintas técnicas y sistemas.

Spärck-Jones y van Rijsbergen (1975) señalaron que el principal problema de las pruebas elaboradas hasta aquel momento era su limitado tamaño (pocos documentos y consultas) y apuntaron la necesidad de nuevas[1] y mayores colecciones así como el modo de generar las listas de documentos relevantes para la posterior evaluación (el conocido método de *pooling*). El trabajo editado por Spärck-Jones (1981) constituye un punto de inflexión al revisar el trabajo realizado hasta finales de los años 1970 y esbozar las líneas a seguir en las décadas posteriores.

A lo largo de los años siguientes comenzaron a desarrollarse conferencias que pretendían, por un lado, ofrecer un marco de experimentación común a fin de poder comparar distintos sistemas y, por otro, crear colecciones de un tamaño similar a las que debería afrontar un sistema real. Así, en 1992 tuvo lugar la primera edición de *TREC –Text REtrieval Conference*, en 1999 la primera de *NTCIR – NII-NACSIS Test Collection for IR Systems* y en 2000 la primera edición de *CLEF – Cross Language Evaluation Forum*.

TREC proporcionó en su primera convocatoria colecciones de entrenamiento y prueba que contenían alrededor de 1GB de texto en lengua inglesa y consistió en dos tareas: consulta y filtrado de información (Harman 1993). En sucesivas ediciones se incluirían nuevas tareas e idiomas, incluyéndose búsquedas bi y multilingües[2].

NTCIR comenzó de manera similar a *TREC* aunque dirigida al idioma japonés y con una colección de 330.000 documentos (Kando 1999). Al igual que *TREC* con el tiempo se incluyeron otros idiomas de interés en el ámbito nipón (p.ej. coreano, chino e inglés) además de tareas de recuperación de información bi y multilingüe y tareas como la extracción de resúmenes automáticos.

---

[1] Poco después Edward Fox (1983) describiría dos nuevas colecciones, *CACM* e *ISI* (o *CISI*), que, sin embargo, continúan la tradición de colecciones "pequeñas" puesto que la mayor contiene sólo 3.204 documentos. Spärck-Jones y Webster (1979) construirían una de las primera colecciones razonablemente grandes, la *NPL*, con casi 11.500 documentos.

[2] Un sistema *IR* multilingüe permite consultar una colección que contiene documentos escritos en distintos idiomas utilizando cualquiera de los mismos en las consultas y obteniendo resultados que incluirán, naturalmente, varios idiomas. Un sistema bilingüe no es más que una simplificación de este caso general.

CLEF, aunque incluye tareas de búsqueda monolingüe, nació teniendo como objetivo la recuperación de información en entornos multilingües debido al contexto europeo. En su primera edición (Peters 2001) se dispuso de colecciones de artículos periodísticos de un año completo en inglés, francés, alemán e italiano y temas de búsqueda en 8 idiomas europeos. Posteriormente, y de modo análogo a los otros dos foros de evaluación, se han ido añadiendo nuevas tareas (p.ej. respuesta de preguntas, búsqueda de imágenes, de documentos sonoros o en la Web) y se ha ampliado el número de idiomas tanto en las colecciones como en los temas de consulta.

En la actualidad las colecciones de prueba existentes para evaluar sistemas *IR* cuentan con cientos de miles de documentos en múltiples idiomas y algunas de las áreas de investigación más activas incorporan aspectos como el multilingüismo, la interactividad en los procesos de búsqueda, la utilización de documentos hipertextuales o la respuesta de preguntas.

4 Utilización de *blindLight* como técnica de recuperación de información[1]

Como se recordará, el autor afirmó como parte de su tesis que la nueva técnica que proponía podía ser empleada como método de recuperación de información. A continuación se describirá el modo en que es posible adaptar *blindLight* a este fin y, posteriormente, se presentarán los resultados obtenidos con la misma al aplicarse sobre colecciones "clásicas" y en un foro de evaluación internacional como es el *CLEF*.

La utilización de *blindLight* como técnica *IR* es muy sencilla, puesto que tanto documentos como consultas son transformados en los correspondientes vectores de *n*-gramas sin ningún tipo de procesamiento previo; en particular, no se realiza *stemming* ni se eliminan las "palabras vacías"[2]. Esto no sólo facilita la aplicación de la técnica a múltiples idiomas sino que, además, las "palabras vacías" contribuyen de manera sustancial al significado de un texto y proporcionan importantes pistas sobre el dominio de conocimiento (Riloff 1995) por lo que no parece adecuado eliminarlas[3].

Por tanto, al igual que en anteriores aplicaciones de esta técnica, lo único necesario es una medida de similitud entre consultas y documentos que estará construida sobre las dos medidas previamente mencionadas Π y P. Como se recordará, P es la relación entre la significatividad total del vector intersección consulta-documento y la del vector documento mientras que Π es la relación entre la significatividad del mismo vector intersección y la del vector consulta. Puesto que el número de *n*-gramas en consultas y documentos son en general muy distintos, los valores de Π y P no son directamente comparables ya que el primero suele ser mucho mayor que el segundo. Por esa razón se han experimentado hasta el momento diversas formas de "normalización" en las distintas medidas de similitud probadas (véase Fig. 91).

---

[1] Este apartado constituye una evolución de las ideas presentadas en Gayo Avello *et al.* (2004b y 2004c).

[2] Lo cierto es que en ninguna de las aplicaciones de *blindLight* se eliminan las palabras vacías por las mismas razones aquí argumentadas.

[3] En las pruebas realizadas se ha comprobado que la eliminación de palabras vacías mejora sustancialmente el rendimiento cuando no se utiliza ningún método que pondere los pesos de los *n*-gramas en función de su distribución en la colección (véase el apartado "Ponderación inter e intradocumental de los *n*-gramas" en la página 145). No obstante, el rendimiento obtenido empleando únicamente dicha ponderación resulta superior al que se alcanza al eliminar palabras vacías y, además, al combinar este tipo de ponderación con la eliminación de palabras vacías la mejora del rendimiento es inapreciable.

$$S_1 = \Pi$$

$$S_2 = \frac{\Pi + norm(\Pi \cdot \mathrm{P})}{2}$$

$$S_3 = \frac{\Pi + \frac{numgrams(doc)}{numgrams(query)}\mathrm{P}}{2}$$

$$S_4 = \frac{numgrams(query \Omega doc)}{numgrams(doc)}\Pi + \frac{numgrams(query \Omega doc)}{numgrams(query)}\mathrm{P}$$

Fig. 91 Medidas de similitud para un sistema *IR* basado en *blindLight*.

query y *doc* son vectores de *n*-gramas que representan una consulta y un documento, respectivamente. *query*Ω*doc* es el vector intersección de ambos vectores. La función *norm* escala los valores que recibe en el intervalo recorrido por Π a fin de hacerlos comparables mientras que la función *numgrams* retorna el número de *n*-gramas (componentes) de un vector.

Una posibilidad muy interesante a desarrollar en el futuro es la utilización de **programación genética** para encontrar medidas de similitud adaptadas a distintos contextos. Después de todo, el número de *n*-gramas en los vectores de consultas, documentos e intersecciones así como los valores Π y P para cada par `(consulta, documento)` son constantes por lo que sería factible obtener dichos datos para una colección estandarizada y emplear los datos de relevancia a modo de "entrenamiento".

La utilización de programación genética para descubrir funciones de ordenación en sistemas *IR* ya ha sido propuesta por Fan, Gordon y Pathak (2004a y 2004b) y permite obtener funciones que mejoran de manera sustancial métodos como *BM25* (Wang *et al.* 2004). No obstante, el autor consideró que esta línea de investigación quedaba fuera del ámbito de este trabajo y este capítulo recogerá tan sólo los resultados obtenidos con las medidas de similitud expuestas arriba.

Así pues, el funcionamiento de un sistema *IR* basado en *blindLight* es, conceptualmente, muy simple:

- Para cada documento de la colección se calcula y almacena un vector de *n*-gramas que lo representa.
- Cuando el sistema recibe una consulta también la representa mediante un vector de *n*-gramas que será comparado con cada vector documento calculando los valores Π y P correspondientes.
- A partir de estos valores es posible calcular una de las anteriores medidas de similitud que se utilizará para ordenar la lista de documentos que satisfacen la consulta. Aquellos documentos más semejantes a la consulta aparecerán antes que otros con una menor similitud.

Ciertamente, una implementación directa de este modo de funcionamiento es muy ineficiente; sin embargo, es posible desarrollar un sistema basado en *blindLight* que emplee técnicas más eficaces (p.ej. un índice invertido de ficheros implementado sobre tablas *hash* en disco).

4.1 *blindLight* como método *CLIR (Cross Language IR)*

Por otro lado, resulta muy sencillo implementar sistemas de recuperación de información multilingües empleando *blindLight* mediante la técnica de "pseudo-traducción" (Gayo Avello *et al.* 2004c). Se emplea el término "pseudo-traducción" puesto que la técnica

no trata de obtener una traducción de las consultas sino vectores que contengan *n*-gramas que aparecerían en unas hipotéticas traducciones.

Para ello se emplea un *corpus* paralelo[1] de los idiomas fuente (*F*) y objeto (*O*) alineado a nivel de sentencias y se procede del modo siguiente (véase Fig. 92). (1) Dada una consulta escrita en el lenguaje fuente, Q_F, se divide en secuencias de palabras de longitud variable (desde una palabra a la consulta completa). (2) Se explora el *corpus F* en busca de sentencias que contengan alguna de dichas secuencias. (3) Cada sentencia (hasta un máximo *k*) encontrada en *F* es reemplazada por su homóloga en el *corpus O*. (4) Para cada una de estas sentencias homólogas en *O* se obtiene un vector de *n*-gramas y todos estos vectores se intersecan empleando el operador **Ω** (descrito en la página 64). (5) Los distintos vectores obtenidos por intersección se mezclan obteniéndose un vector consulta pseudo-traducido.

Puesto que todas las sentencias homólogas encontradas en *O* contienen presumiblemente la traducción del mismo conjunto de palabras del idioma *F* parece razonable suponer que la intersección **Ω** de los correspondientes vectores contendrá un conjunto de *n*-gramas "traducidos". Por otro lado, aquellas palabras de la consulta que no aparecen en el *corpus* fuente son incluidas directamente en la pseudo-traducción. En teoría, este proceso daría lugar a vectores consulta similares a los que se obtendrían a partir de traducciones reales de las consultas originales.

Este método de pseudo-traducción fue puesto en práctica durante la participación del autor en *CLEF 2004* (Gayo Avello *et al.* 2004c). Puesto que en dicha campaña todos los idiomas objeto de estudio (a excepción del ruso) eran lenguas de la Unión Europea se utilizó como *corpus* paralelo el denominado *Europarl*[2] (Koehn) obteniéndose resultados muy interesantes. Así, para comprobar la "calidad" de las pseudo-traducciones se compararon los vectores obtenidos al pseudo-traducir las consultas *CLEF* de castellano a inglés con los vectores correspondientes a las consultas originalmente escritas en inglés resultando que, en promedio, el 38,59% de los *n*-gramas de las pseudo-traducciones aparecen en las traducciones reales y el 28,31% de los *n*-gramas de estas últimas se encuentran en las primeras. No parece necesario decir que este rendimiento debe mejorarse.

Es preciso señalar que esta técnica tiene cierta relación con las propuestas por Pirkola *et al.* (2002) para encontrar palabras equivalentes en distintos idiomas que difieren en su grafía[3] y por McNamee y Mayfield (2003) para "traducir" *n*-gramas (véase Fig. 93). La diferencia entre tales técnicas y la propuesta por el autor es que éste no pretende obtener traducciones ni para palabras ni para *n*-gramas individuales sino generar a partir de un vector de *n*-gramas correspondiente a un texto en un idioma fuente otro vector que contenga aquellos *n*-gramas que con mayor probabilidad formarían parte de una traducción real a un idioma objeto, vector que podría ser enviado directamente como consulta a un sistema *IR* basado en *blindLight*.

Por otro lado, aunque existen similitudes, la técnica de pseudo-traducción aquí presentada es mucho más sencilla que la de traducción de *n*-gramas individuales de McNamee y Mayfield y, al tiempo, ofrece las mismas ventajas que ésta frente a métodos de

---

[1] Un *corpus* paralelo es una colección de textos traducidos a varios idiomas además del original (*EAGLES* 1996).

[2] *European Parliament Proceedings Parallel Corpus 1996-2003*.

[3] Por ejemplo, `Rwanda` y `Ruanda`, `Chechnya` y `Tsetshenia` o `pharmacology` y `farmakologian` (Pirkola *et al.* 2002).

traducción basados en diccionarios. El funcionamiento de ambas técnicas puede compararse en Fig. 92 y Fig. 93.

(1) Consulta original en el idioma F (castellano):

```
Encontrar documentos en los que se habla de las discusiones sobre la reforma de
las instituciones financieras y, en particular, del Banco Mundial y del FMI
durante la cumbre de los G7 que se celebró en Halifax en 1995.
```

(2) Fragmentos de la consulta a buscar en el corpus F:

```
Encontrar
Encontrar documentos
Encontrar documentos en
…
instituciones
instituciones financieras
…
```

(3) Sentencias del corpus F que contienen el fragmento anterior:

(1315) …mantiene excelentes relaciones con las instituciones financieras internacionales…
(5865) …el fortalecimiento de las instituciones financieras internacionales…
(6145) …La Comisión deberá estudiar un mecanismo transparente para que las instituciones financieras europeas…

(4) Sentencias homólogas en el corpus O (inglés):

(1315) …has excellent relationships with the international financial institutions…
(5865) …strengthening international financial institutions…
(6145) …The Commission will have to look at a transparent mechanism so that the European financial institutions…

(5) Intersección Ω de las sentencias de O homólogas de `instituciones financieras`:

```
{' fi', ' in', 'al ', 'anc', 'cia', 'fin', 'ial', 'ina', 'ins', 'ion', 'itu',
 'l i', 'nan', 'nci', 'nst', 'ons', 'sti', 'the', 'tio', 'tit', 'tut', 'uti'}
```

(6) Vector final correspondiente a la pseudo-traducción de la consulta:

```
{'Bank', 'Worl', 'ld B', ..., 'Hali', 'ifax', ..., 'cumb', 'mbre', ..., 'ssio',
'ussi', 'ards', 'ax e', ' IMF', ..., 'FMI ', 'bre ', 'n Ha', ..., ' G7 ', ' by ',
'the ', ..., 't th', 'n th'}
```

Solapamiento entre la hipotética traducción y el vector pseudo-traducido:

```
Find documents about discussions on the reform of financial institutions, and in
particular the World Bank and the IMF, at the G7 summit that took place in
Halifax in 1995
```

Términos del vector pseudo-traducido que no se corresponden con la traducción hipotética:

```
{' by ', ' cum', ' en ', ' FMI', ' la ', ' la r', ' los', ' oth', ' pos', ' pro',
' rar ', ' to ', ' Uni', 'a re', 'annu', 'ards', 'atin', 'atio', 'ax e', 'bili',
'bre ', 'cont', 'cumb', 'e su', 'e to', 'en H', 'en p', 'Enco', 'ere ', 'FMI ',
'her ', 'ibil', 'ilit', 'in t', 'ing ', 'ion ', 'los ', 'mbre', 'ncon', 'nnua',
'nt t', 'ntra', 'nual', 'ontr', 'orma', 'ossi', 'othe', 'ould', 'poss', 'rds ',
'ring', 'rma ', 's of', 'sibi', 'ssib', 'ted ', 'ther', 'trar', 'tten', 'ual ',
'uld ', 'umbr', 'x en'}
```

Fig. 92 Proceso de pseudo-traducción de una consulta.

Los pasos 1 al 6 muestran el modo en que se lleva a cabo el proceso de pseudo-traducción de una consulta de un idioma F (castellano) a un idioma O (inglés). (4) y (5) muestran los *n*-gramas comunes a las sentencias homólogas. Se incluye además la traducción real de la consulta (que no se emplea en el proceso de pseudo-traducción) junto con el solapamiento entre la misma y la pseudo-traducción. Por último se presentan aquellos *n*-gramas incluidos en el vector incorrectamente pues no pertenecen a la traducción real.

Por último, esta técnica de pseudo-traducción puede resultar un campo de trabajo prometedor puesto que, por un lado, no persigue obtener traducciones de textos destinadas a "consumo humano" lo cual la simplifica enormemente y, por otro, existen toda una serie

de aspectos en los que es posible introducir mejoras. Por ejemplo, la intersección de los vectores de sentencias homólogas puede resultar excesivamente simplista, además, deberían analizarse diferentes *corpora* paralelos[1] así como la posibilidad de emplear *corpora* comparables[2].

```
communist party

_comm  commu  ommun  mmuni  munis  unist  nist_  ist_p  st_pa  t_par  _part  party  arty_
mmuna  munau  munau  munau  munis  unist  unist  ist_p  l_re_  rtie_  _part  rtie_  rtie_

parti communiste
```

Fig. 93 Método de traducción de *n*-gramas de McNamee y Mayfield (2003).

McNamee y Mayfield utilizan una técnica que permite asignar a cada *n*-grama de un idioma fuente (en este caso inglés) un único *n*-grama en un idioma objeto (francés). En este ejemplo proporcionado por sus autores se muestra la "traducción" al francés de los *n*-gramas de `communist party` junto y el solapamiento de los *n*-gramas traducidos y una traducción real.

4.2 Ponderación inter e intradocumental de los *n*-gramas

Ya se ha señalado con anterioridad que la relevancia de un término está estrechamente relacionada con el número de documentos que lo contienen: cuanto mayor es este número menos relevante es el término y, viceversa, la relevancia es mayor en el caso de términos poco comunes. Spärck-Jones (1972) fue la primera en presentar una técnica de ponderación, *idf*, basada en esta hipótesis cuya solidez se ha comprobado no sólo desde un punto de vista empírico sino también teórico (Robertson 2004).

Por lo que respecta al modelo *IR* basado en *blindLight* aún no se ha presentado en este trabajo ningún método de ponderación análogo: cada *n*-grama de un documento tiene un peso obtenido exclusivamente a partir del propio documento sin utilizar ninguna información sobre la distribución de los distintos *n*-gramas en la colección. La ventaja de este enfoque radica en que la colección puede crecer indefinidamente sin necesidad de re-indexar puesto que este proceso se realiza una única vez por documento en el momento en que se añade éste a la colección.

No obstante, a pesar de esta ventaja se está desaprovechando información presente en la colección. Por ello resulta importante analizar el modo de obtener tal información así como su influencia en el rendimiento del sistema a fin de valorar la relación coste-beneficio de su aplicación. Así pues, este apartado presentará un método similar a *idf* y proporcionará más detalles acerca de la ponderación de los *n*-gramas dentro de cada documento.

Para cada documento, D_i, de una colección es posible obtener un vector que asocie a cada *n*-grama del texto su significatividad. Es posible entonces obtener un valor promedio[3]

---

[1] Por ejemplo, *OPUS* (`http://logos.uio.no/opus`) o *MULTEXT-East* (`http://nl.ijs.si/ME`).

[2] Un *corpus* comparable es aquel que contiene textos similares en más de un lenguaje o variedad (*EAGLES* 1996). Reinhard Rapp (1999) señala que dicha similitud viene marcada por un dominio de conocimiento común a todos los textos. En general se considera que un *corpus* comparable consta de varias colecciones de documentos de tamaño y temática similares y que proporcionan aproximadamente el mismo número de términos para cada uno de los idiomas implicados. El gran atractivo de los *corpora* comparables es que son relativamente sencillos de construir en comparación con los paralelos. Así, por ejemplo, en el marco del *CLEF* muchos autores han empleado las propias colecciones de noticias como *corpora* comparables (Rogati y Yang 2001), (Cancedda *et al.* 2003) o (Peinado *et al.* 2004).

[3] Estos valores estadísticos son obtenidos para cada *n*-grama dentro de la colección, es decir, no se calcula el valor promedio de la significatividad de un *n*-grama dentro de los documentos que lo contienen sino en todos los documentos aun cuando no lo incluyan.

para la significatividad de cada *n*-grama que aparezca en la colección así como su desviación típica y su coeficiente de variación[1]. Parece razonable suponer que aquellos *n*-gramas más comunes (p.ej. los correspondientes a palabras vacías) no sólo aparecerán en muchos documentos sino que lo harán con significatividades similares por lo que su coeficiente de variación será reducido mientras que los *n*-gramas más raros presentarán un coeficiente de variación mayor[2].

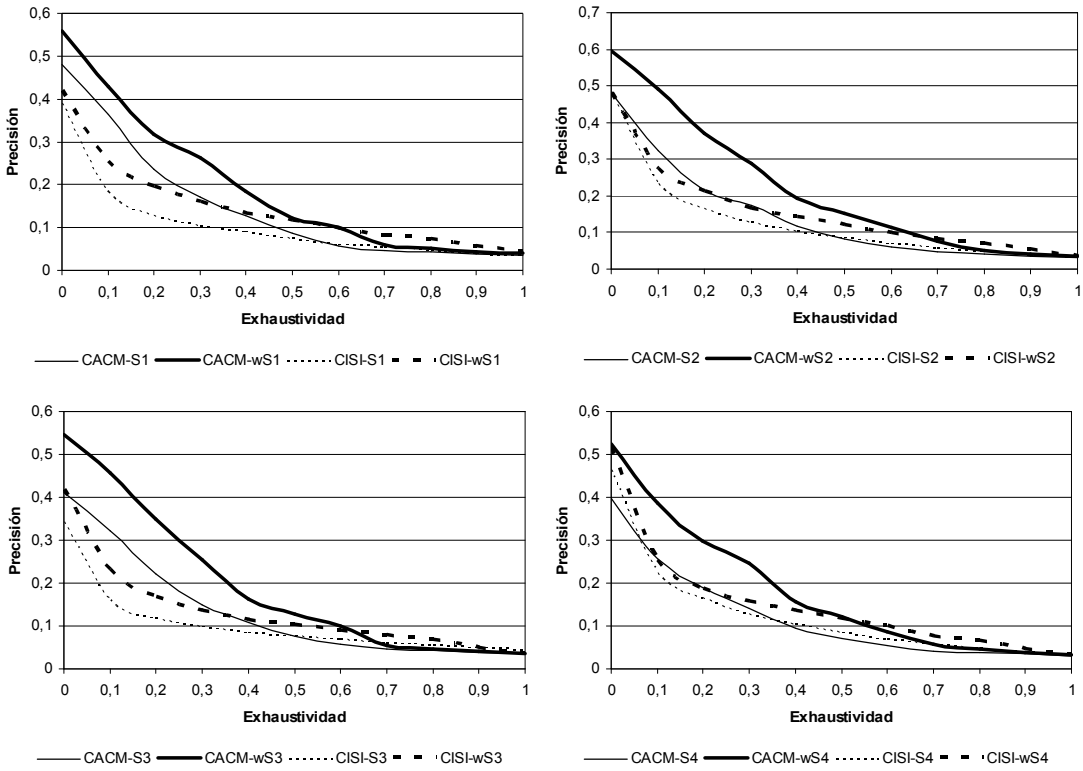


Fig. 94 Influencia de la ponderación interdocumental basada en el coeficiente de variación.

Se presentan los gráficos precisión-exhaustividad para los resultados obtenidos con las colecciones *CACM* y *CISI* empleando las cuatro medidas de similitud presentadas anteriormente. Las medidas no ponderadas están etiquetadas como *sN* (con *N* variando entre 1 y 4) y emplean trazo fino; las medidas ponderadas usan como etiquetas *wSN* y se muestran con trazo grueso. Como era de esperar, en todos los casos la introducción de la ponderación interdocumental supone una mejora sustancial en el rendimiento.

Así pues, al emplear *blindLight* como un sistema *IR*, documentos y consultas se representan mediante vectores de *n*-gramas que tienen asignado como peso el producto de la significatividad del *n*-grama en el documento por su coeficiente de variación en la colección. Como era de suponer, la utilización de información sobre distribución de los *n*-gramas en la colección supone una mejora del rendimiento muy sustancial (véase Fig. 94). Por otro lado, se han estudiado otras formas de ponderación[3], incluyendo *idf*, y en todos los casos el coeficiente de variación ha resultado notablemente superior (véase Fig. 95).

---

[1] El coeficiente de variación es el cociente de la desviación típica entre la media.

[2] El valor máximo que puede alcanzar el coeficiente de variación es $\sqrt{N-1}$ donde N es el tamaño de la colección.

[3] Además de *idf* se ha experimentado con una versión normalizada del coeficiente de variación en el intervalo [0, 1] y con el logaritmo del coeficiente de variación.

Por otro lado, aun cuando a lo largo de este trabajo se ha venido empleando la información mutua para determinar la significatividad, y por tanto los pesos, de los *n*-gramas dentro de cada documento existen otras posibilidades. Por ejemplo, la probabilidad condicional simétrica, los coeficientes Dice y ϕ^2 (Ferreira da Silva y Pereira Lopes 1999) o la ganancia de información (véase Fig. 96).

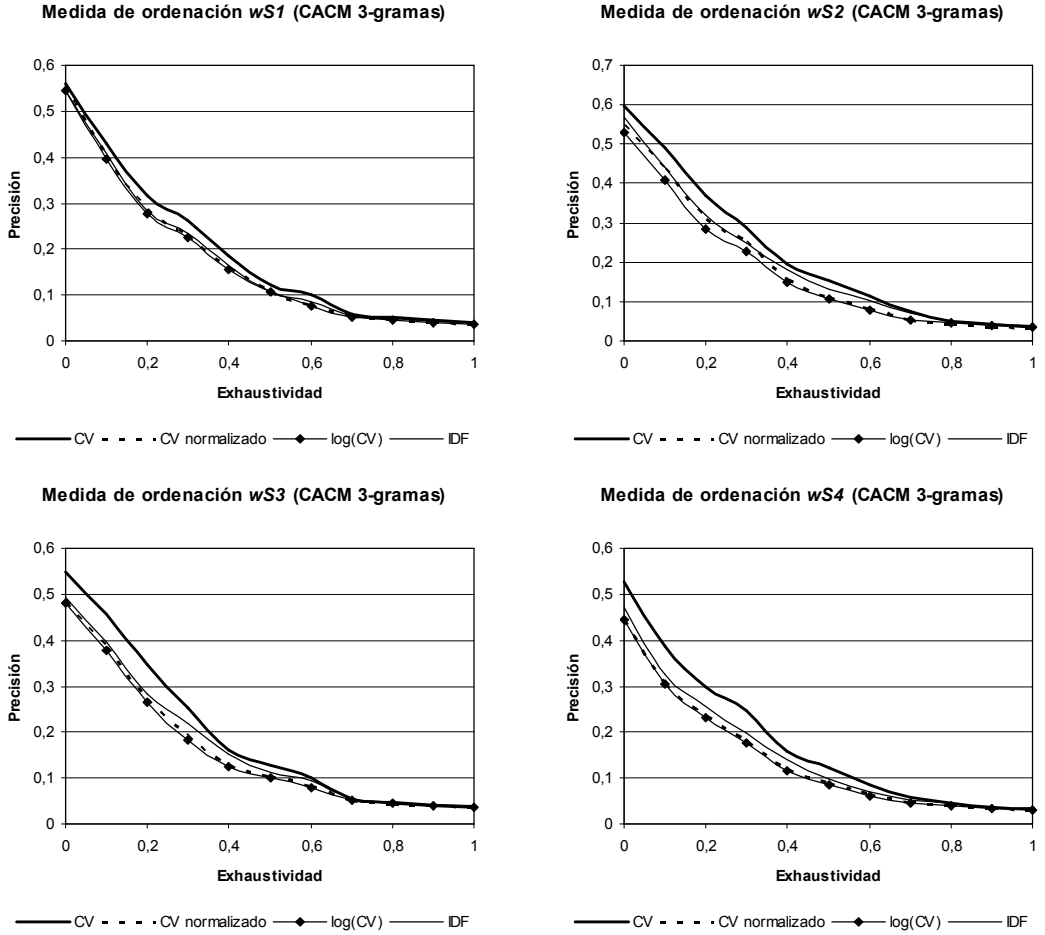


Fig. 95 Distintos métodos de ponderación interdocumental.

Al compararlo con otros métodos (incluido *idf*) el coeficiente de variación resultó el más eficaz siendo las diferencias sustanciales en la práctica totalidad de los casos.

Como se recordará, Ferreira da Silva y Pereira Lopes (1999) desarrollaron una técnica que permitía generalizar una serie de estadísticos para *n*-gramas de longitud arbitraria (véase página 61). Dichos autores utilizaron esas medidas para determinar el grado de "pegajosidad" de *n*-gramas de palabras facilitando así la extracción de términos multipalabra[1]. El autor de esta disertación propuso aplicar la misma técnica a *n*-gramas de caracteres a fin de determinar su grado de significatividad dentro de un texto. En las aplicaciones descritas hasta el momento (clasificación y categorización) se ha empleado la información mutua y en el prototipo participante en *CLEF'04* la probabilidad condicional simétrica.

[1] Simplificando enormemente, cuanto más "pegajoso" resulta un *n*-grama mayor es la probabilidad de que sea un término multi-palabra.

$$Avp = \frac{1}{n-1}\sum_{i=1}^{i=n-1} p(w_1...w_i)\cdot p(w_{i+1}...w_n)$$

$$SI\_f((w_1...w_n)) = \log(\frac{p(w_1...w_n)}{Avp}) \quad (1)$$

$$SCP\_f((w_1...w_n)) = \frac{p(w_1...w_n)^2}{Avp} \quad (2)$$

$$Avx = \frac{1}{n-1}\cdot\sum_{i=1}^{i=n-1} f(w_1...w_i) \qquad Avy = \frac{1}{n-1}\cdot\sum_{i=2}^{i=n} f(w_i...w_n)$$

$$\phi^2\_f((w_1...w_n)) = \frac{[f(w_1...w_n)\cdot N - Avp]^2}{Avp\cdot(N-Avx)\cdot(N-Avy)} \quad (3)$$

$$Dice((w_1...w_n)) = \frac{2\cdot f(w_1...w_n)}{Avx + Avy} \quad (4)$$

$$Infogain((w_1...w_n)) = \frac{1}{n-1}\sum_{i=1}^{i=n-1} p(w_1...w_i)\cdot\log\frac{1}{p(w_1...w_i)} + p(w_{i+1}...w_n)\cdot\log\frac{1}{p(w_{i+1}...w_n)} \quad (5)$$

Fig. 96 Estadísticos para la ponderación de *n*-gramas dentro de un documento.

(w1…wn) es un *n*-grama, *(w1…wi)* y *(wi+1…wn)* son fragmentos consecutivos del mismo (p.ej. para el *n*-grama *'info'* se tendría <*'i'*, *'nfo'*>, <*'in'*, *'fo'*> y <*'inf'*, *'o'*>). *p((w1…wn))* es la probabilidad del *n*-grama *(w1…wn)* en el texto, *p((w1…wi))* es la probabilidad de que un *n*-grama comience con los caracteres *(w1…wi)* y *p(wi+1…wn)* de que termine en *(wi+1…wn)*. *f((w1…wn))*, *f((w1…wi))* y *f((wi+1…wn))* son frecuencias absolutas. *N* es el número de *n*-gramas distintos en el documento. Los estadísticos del (1) al (4) fueron propuestos por Ferreira da Silva y Pereira Lopes (1999) y el quinto por el autor.

A continuación se presentan algunos resultados sobre la influencia del "estadístico de ponderación intradocumental" (véase Fig. 97, Fig. 98, Fig. 99 y Fig. 100); sin embargo, deben considerarse preliminares y es necesario un estudio detallado que combine distintos (1) estadísticos, (2) tamaños de *n*-grama y (3) medidas de similitud (esto es, combinaciones de Π y P). No obstante, puesto que la nueva técnica propuesta no está vinculada a ninguna medida de la significatividad en particular y tan sólo se ha señalado la existencia y viabilidad de varias de tales medidas, un análisis exhaustivo de la eficacia de las mismas queda fuera de los objetivos de este trabajo.

A la vista de semejantes datos no es posible llegar a ninguna conclusión definitiva puesto que, aunque existen diferencias de rendimiento sustanciales, no hay ningun estadístico que resulte claramente superior al resto en todos los casos (es decir, para todas las medidas de similitud). No obstante, la información mutua parece la opción más acertada, especialmente si se combina con el método de ponderación basado en el coeficiente de variación del peso de los *n*-gramas en la colección.

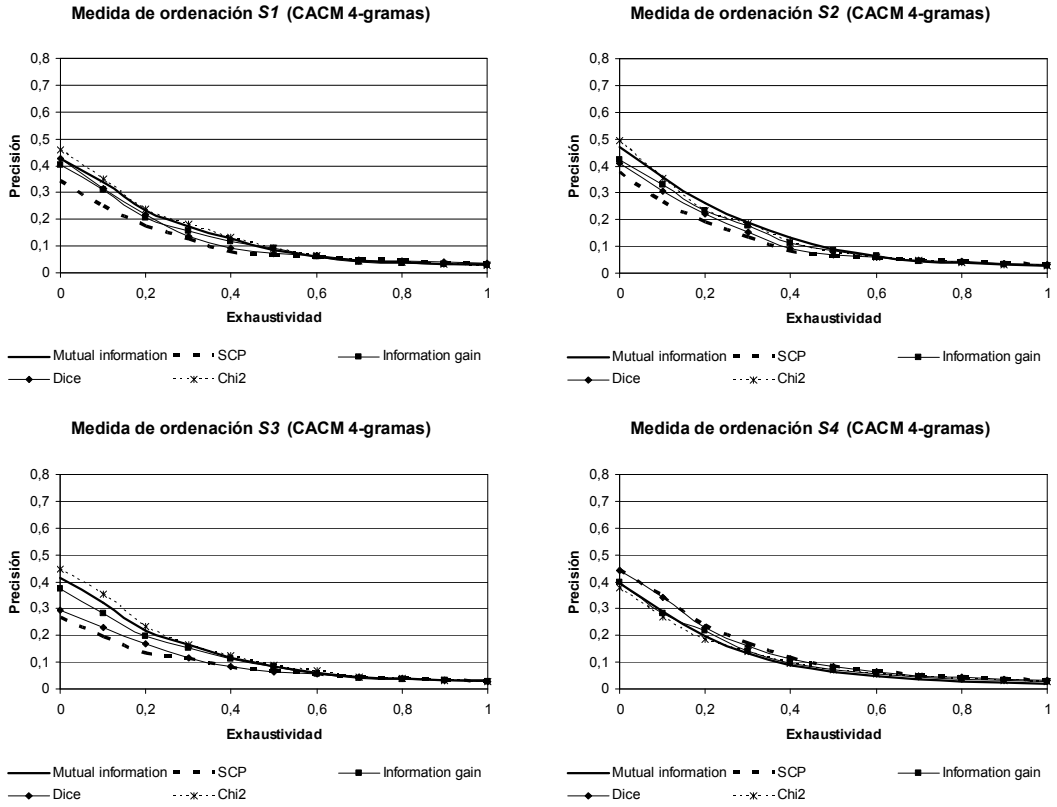


Fig. 97 Rendimiento de los distintos estadísticos para el cálculo del peso de los *n*-gramas.

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	**0,1447**	-3,9%
SCP	0,1142	**-24,1%**
Information gain	0,1352	**-10,2%**
Dice	0,1355	**-10,0%**
ϕ^2	**0,1506**	

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	**0,1553**	
SCP	0,1205	**-22,4%**
Information gain	0,1426	**-8,2%**
Dice	0,1343	**-13,5%**
ϕ^2	**0,1520**	-2,1%

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	0,1389	**-6,1%**
SCP	0,0956	**-35,4%**
Information gain	0,1271	**-14,1%**
Dice	0,1058	**-28,5%**
ϕ^2	**0,1480**	

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	0,1200	**-18,9%**
SCP	**0,1479**	
Information gain	0,1281	**-13,4%**
Dice	**0,1450**	-1,9%
ϕ^2	0,1223	**-17,3%**

Fig. 98 Rendimiento de los distintos estadísticos para la ponderación de *n*-gramas.

(De izquierda a derecha y de arriba abajo, S1, S2, S3 y S4). Como se puede comprobar las diferencias de rendimiento son notables en todos los casos y no puede asegurarse que un estadístico sea claramente superior al resto puesto que los resultados dependen enormemente de la medida de ordenación de resultados utilizada. No obstante, parece que el coeficiente ϕ^2 y la información mutua son los estadísticos que ofrecen de manera consistente los mejores resultados para la mayor parte de funciones de ordenación.

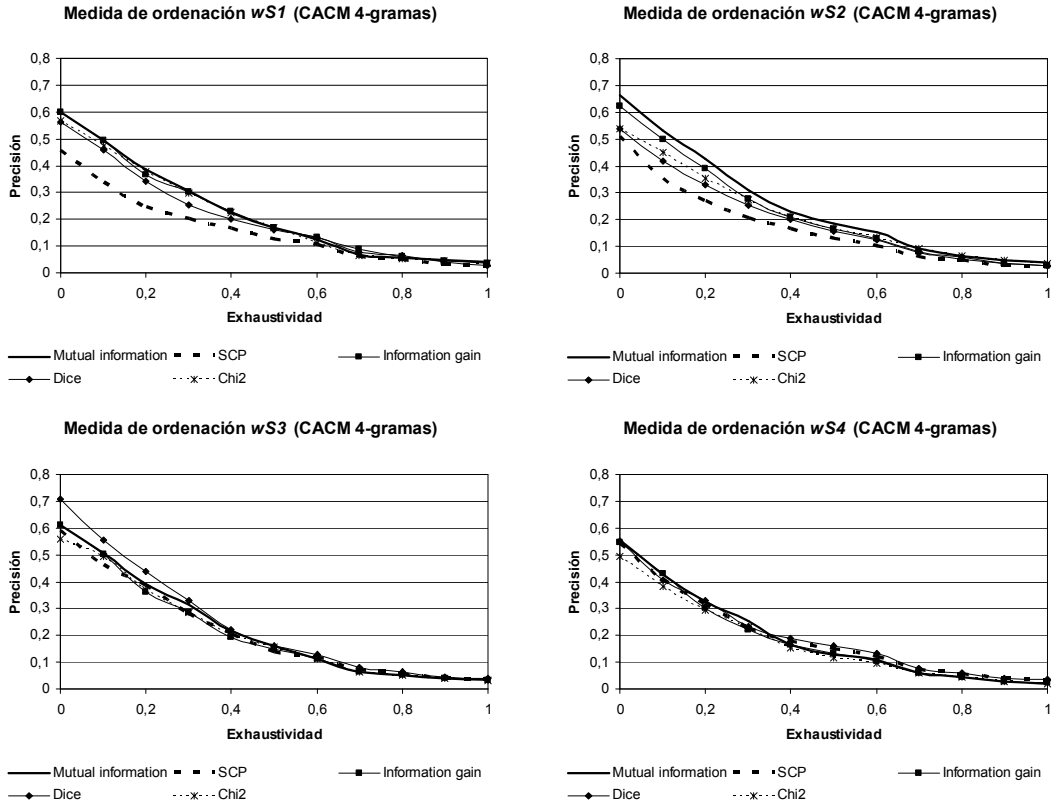


Fig. 99 Rendimiento de los distintos estadísticos para el cálculo del peso de los *n*-gramas (ii).

En este caso se ha utilizado el coeficiente de variación del peso de los *n*-gramas para ponderar los vectores de documentos y consultas.

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	**0,2285**	-0,4%
SCP	0,1636	**-28,7%**
Information gain	**0,2293**	
Dice	0,2108	**-8,1%**
ϕ^2	0,2199	**-4,1%**

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	**0,2492**	
SCP	0,1729	**-30,6%**
Information gain	0,2267	**-9,0%**
Dice	0,2026	**-18,7%**
ϕ^2	0,2150	**-13,7%**

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	0,2275	**-9,7%**
SCP	0,2169	**-14,0%**
Information gain	0,2188	**-13,2%**
Dice	**0,2521**	
ϕ^2	0,2149	**-14,7%**

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	**0,1919**	-4,6%
SCP	**0,1946**	-3,2%
Information gain	0,1869	**-7,0%**
Dice	**0,2021**	
ϕ^2	0,1744	**-13,3%**

Fig. 100 Rendimiento para los distintos estadísticos para la ponderación de *n*-gramas (ii).

(De izquierda a derecha y de arriba abajo, wS1, wS2, wS3 y wS4). En este caso se ha empleado la técnica de ponderación de *n*-gramas basada en el coeficiente de variación. De nuevo hay diferencias sustanciales en los rendimientos obtenidos. Sin embargo, la información mutua parece perfilarse como uno de los estadísticos más adecuados a la hora de determinar el peso de los *n*-gramas dentro de cada documento.

4.3 Influencia del tamaño de *n*-grama utilizado

El tamaño de *n*-grama utilizado al representar documentos y consultas tiene, como era de esperar, un impacto en el rendimiento aunque menor de lo previsto, dependiendo de la naturaleza de la colección y con resultados "extraños". Así, tanto en la colección *CACM*

como en la colección *CISI* hay un aumento en el rendimiento al pasar de 3-gramas a 4-gramas; sin embargo, mientras que en la colección *CACM* el cambio es sustancial en la colección *CISI* es inapreciable. En cambio, si se comparan los resultados obtenidos en ambos casos al emplear 5-gramas y 3-gramas la mejoría es apreciable (cercana al 10%). Por otro lado, las diferencias entre el uso de 4-gramas y 5-gramas son pequeñas: apreciables en el caso de la colección *CISI* e inapreciables en el caso de *CACM*, aunque, extrañamente, los resultados obtenidos para esta colección empleando 5-gramas son ligeramente peores que utilizando 4-gramas.

En definitiva, el tamaño de *n*-grama tiene una influencia en los resultados obtenidos por el sistema difíciles de evaluar *a priori* en tanto en cuanto parecen venir determinados por la naturaleza de los textos de la colección. Habida cuenta de este hecho, de las distintas posibilidades que se han descrito para la obtención de los pesos de los *n*-gramas en cada documento, así como de las diferentes combinaciones de las medidas Π y P en una única función de ordenación, se abre una línea de trabajo destinada a evaluar de manera sistemática el rendimiento de las distintas configuraciones de *blindLight* al trabajar sobre colecciones e idiomas diversos.

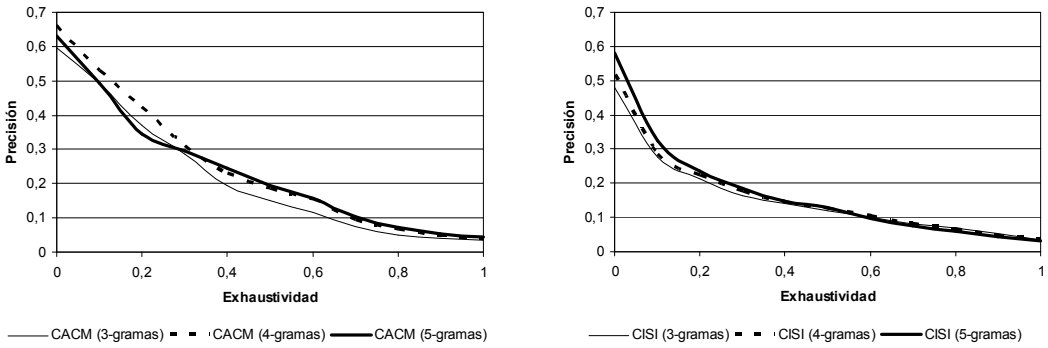


Fig. 101 Curvas *P-R* para las colecciones *CACM* y *CISI* usando distintos tamaños de *n*-grama.

5 Resultados obtenidos por *blindLight*. Comparación con otras técnicas

A fin de estudiar la viabilidad de *blindLight* como técnica de recuperación de información se implementó un prototipo que participó en *CLEF'04* (Gayo Avello *et al.* 2004c). Previamente se realizaron unas pruebas con dos colecciones de menor tamaño, *CACM* y *CISI* (Fox 1983), de las cuales se han presentado algunos resultados a lo largo de los apartados anteriores.

```
.I 47
.W
The use of Bayesian decision models to optimize information
retrieval system performance.  This includes stopping rules to
determine when a user should cease scanning the output of a
retrieval search.
.N 47. Donald Kraft
```

Fig. 102 Una consulta para la colección *CACM*.

La primera colección consiste en un conjunto de títulos y resúmenes de artículos publicados en la revista *Communications of the ACM* entre 1958 y 1979. En total consta de 3204 documentos y 64 consultas (véase Fig. 102 y Fig. 103) junto con los correspondientes "juicios de relevancia". La colección *CISI* consta de 1460 documentos (también resúmenes) y 112 consultas y se proporciona en un formato análogo al de la colección *CACM*. Al preparar los vectores de *n*-gramas para los documentos de ambas colecciones se utilizó el

título, contenido y autor del documento pero no las referencias a otros documentos. Por lo que respecta a las consultas se empleó únicamente el texto de la consulta y nunca el autor de la misma.

```
.I 1457
.T
Data Manipulation and Programming Problems
in Automatic Information Retrieval
.W
Automatic information retrieval programs require the
manipulation of a variety of different data structures,
including linear text, sparse matrices, and tree or list
structures.  The main data manipulations to be performed in
automatic information
systems are first briefly reviewed.  A variety of data
representations which have been used to describe structured
information are then examined, and the characteristics of
various processing languages are outlined in the light of the
procedures requiring implementation.  Advantages of these
programming languages for the retrieval application are
examined, and suggestions are made for the design of programming
facilities to aid in information retrieval.
.B
CACM March, 1966
.A
Salton, G.
.N
CA660315 JB March 3, 1978  11:35 AM
.X
1457       4       1457
1236       5       1457
1457       5       1457
1457       5       1457
1457       5       1457
```

Fig. 103 Un documento de la colección *CACM*.

A la vista de los resultados (véase Tabla 22) es innegable que la actual implementación de *blindLight* como técnica de recuperación de información aún está lejos de proporcionar resultados próximos a los de modelos ya consolidados como el vectorial (Kolda y O'Leary 1998) (Crestani y van Rijsbergen 1998) (Carpineto y Romano 2000) (Tombros *et al.* 2002) (Billhardt *et al.* 2003) o el probabilístico (Crestani y van Rijsbergen 1998). No obstante, su rendimiento es semejante al de técnicas como el indexado mediante semántica latente (Kolda y O'Leary 1998) y puesto que existen diversos aspectos de la propuesta que todavía requieren un análisis exhaustivo es previsible obtener un mejor rendimiento en el futuro.

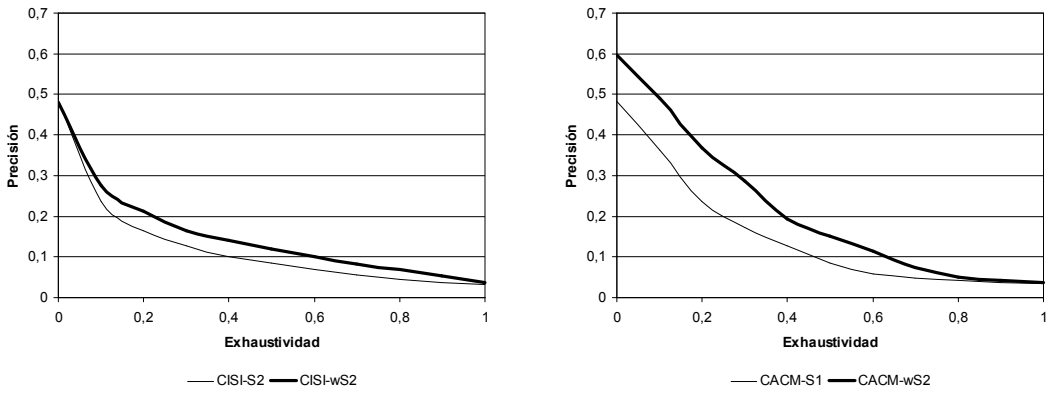


Fig. 104 Resultados obtenidos por *blindLight* sobre las colecciones *CISI* y *CACM*.

	CACM				CISI			
	Precisión media interpolada 11 pt.		Precisión media no interpolada		Precisión media interpolada 11 pt.		Precisión media no interpolada	
blindLight *(wS2, 4-gramas)*	**0,249**		**0,233**		**0,174**		**0,154**	
Modelo vectorial								
(Kolda y O'Leary 1998)	-		-		0,184	5,75%	-	
RbJP (Crestani y van Rijsbergen 1998)	0,271	8,84%	-		-		-	
(Carpineto y Romano 2000)	0,340	36,55%	0,320	37,34%	-		-	
(Tombros *et al.* 2002)	0,378	51,81%	-		0,195	12,07%	-	
(Billhardt *et al.* 2003)	-		0,332	42,49%	-		0,237	53,90%
Hierarchical Clustering (Carpineto y Romano 2000)	0,257	3,21%	0,231	-0,86%	-		-	
Concept Lattice (Carpineto y Romano 2000)	0,281	12,85%	0,253	8,58%	-		-	
Programación genética + CVM[1] (Billhardt *et al.* 2003)	-		0,375	60,94%	-		0,258	67,53%
Modelo probabilístico								
RbLI (Crestani y van Rijsbergen 1998)	0,332	33,33%	-		-		-	
RbCP (Crestani y van Rijsbergen 1998)	0,371	49,00%	-		-		-	
RbGLI (Crestani y van Rijsbergen 1998)	0,428	71,89%	-		-		-	
Semántica Latente								
SDD (Kolda y O'Leary 1998)	-		-		0,181	4,02%	-	
SVD (Kolda y O'Leary 1998)	-		-		0,179	2,87%	-	

Tabla 22. Comparación del rendimiento de *blindLight* en relación con otras técnicas *IR*.

Los resultados obtenidos por *blindLight* aún son sustancialmente inferiores a los alcanzados por modelos como el vectorial o el probabilístico, aunque comparables a los proporcionados por otras técnicas como las de *clustering* jerárquico o semántica latente.

Además, *blindLight* participó en la edición de 2004 del *CLEF* en dos tareas: recuperación de información monolingüe en la colección de documentos escritos en ruso y recuperación bilingüe consultando en castellano la colección de textos escritos en inglés[2]. En el primer caso el prototipo retornó 72 de los 123 documentos relevantes con una precisión media de 0,1433. En la búsqueda bilingüe obtuvo 145 de los 3750 documentos relevantes con una precisión de 0,0644.

5 temas con mejores resultados (ES-EN)		5 temas con mejores resultados (RU)	
218	Andreotti and the Mafia	230	Atlantis-Mir Docking
248	Macedonia Name Dispute	209	Tour de France Winner
202	Nick Leeson's Arrest	210	Nobel Peace Prize Candidates
224	Woman solos Everest	211	Peru-Ecuador Border Conflict
205	Tamil Suicide Attacks	202	Nick Leeson's Arrest
5 temas con peores resultados (ES-EN)		**5 temas con peores resultados (RU)**	
212	Sportswomen and Dopping	227	Altai Ice Maiden
235	Seal-hunting	203	East Timor Guerrillas
241	New political parties	207	Fireworks Injuries
214	Multi-billionaires	228	Prehistorical art
216	Glue-sniffing Youngsters	250	Rabies in Humans

Tabla 23. Temas con los mejores y peores resultados para las tareas monolingüe y bilingüe.

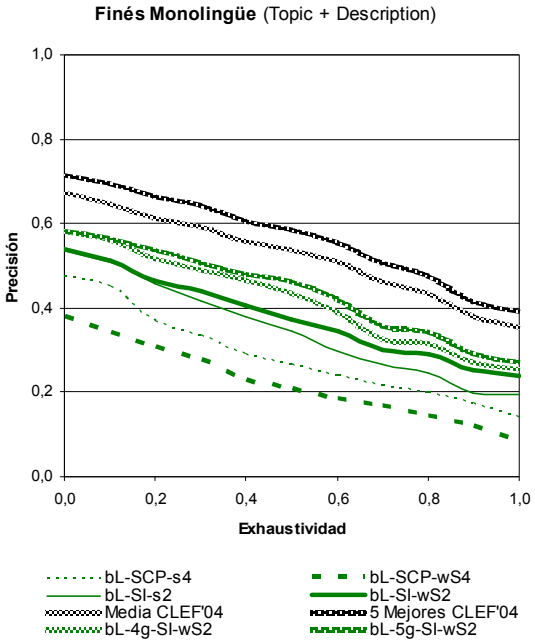
Se han definido los "mejores" como aquellas consultas con la precisión más alta para los 5 primeros documentos retornados y "peores" son las que no retornaron ningún resultado relevante (cuanto mayor era el número de documentos relevantes en la colección peor la consulta). Como se puede ver las consultas relativas a personas, lugares y/o eventos son las que obtienen mejores resultados empleando *blindLight IR* mientras que las consultas abiertas aún no son manejadas de manera adecuada.

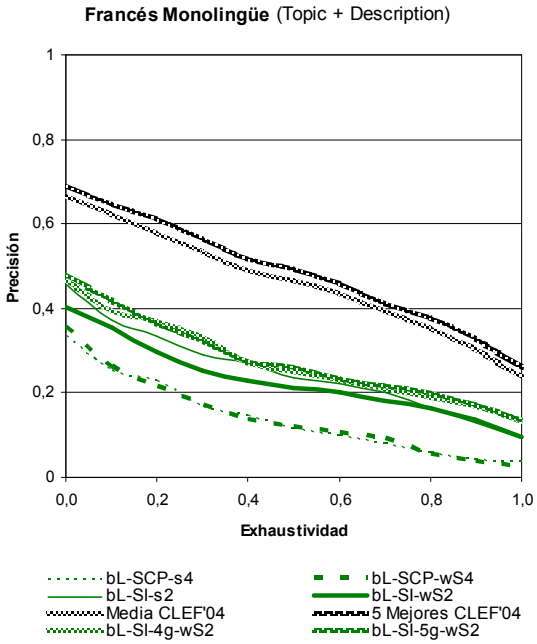
Tales resultados distan mucho de ser buenos pero aun así el autor los consideró alentadores en primer lugar por tratarse de la primera participación en el *CLEF* y en segundo lugar porque aunque el comportamiento promedio es bastante pobre es posible

[1] *Context Vector Model.*

[2] Las consultas escritas originalmente en castellano se pseudo-traducían al inglés y estos vectores de *n*-gramas eran utilizados para consultar la colección de documentos en dicho idioma.

determinar qué clase de temas son los que obtienen peores resultados (véase Tabla 23) señalando una futura línea de trabajo. Hay que señalar, además, que el prototipo participante no empleaba ponderación interdocumental y que el sistema de pseudo-traducción aún está en una fase incipiente todo lo cual influyó sin duda de manera negativa en el rendimiento del sistema en la búsqueda bilingüe.





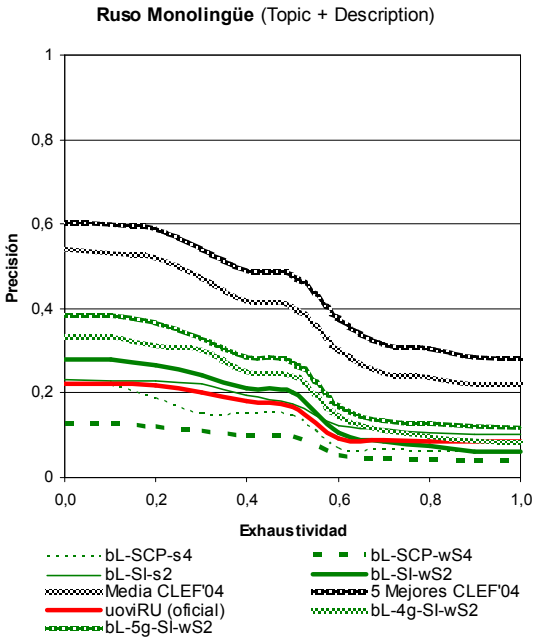


Fig. 105 Pruebas no oficiales de *blindLight* como sistema *IR* monolingüe para finés, francés y ruso.

Las consultas para cada idioma se construyeron automáticamente a partir del título y descripción de los distintos temas (véase Fig. 82). El tamaño de *n*-grama utilizado fue 3 (aunque también se presentan resultados para 4- y 5-gramas) y se experimentó con la probabilidad condicional simétrica y la información mútua como métodos de ponderación intradocumental y con el coeficiente de variación como método de ponderación interdocumental. Para cada combinación de métodos de ponderación se muestra la medida que obtuvo mejores resultados así como los resultados medios de los 5 mejores participantes en *CLEF'04* y los resultados medios de la campaña. En el caso de la prueba en ruso se presenta también el resultado oficial obtenido por el prototipo de *blindLight* (en rojo). Dicho prototipo usaba 3-gramas, *SCP*, la medida de similitud S_2 y no empleaba ponderación interdocumental. Por otro lado, mientras que para el prototipo empleado en *CLEF'04* se transliteró el texto escrito en ruso, el utilizado en estas pruebas puede operar directamente sobre texto *Unicode*. En el caso de finés y francés no hay resultados oficiales pues, aunque inscrito, el autor no obtuvo a tiempo los resultados para su envío a la organización.

Una vez finalizado *CLEF'04* se llevó a cabo una serie de pruebas "no oficiales[1]" de recuperación monolingüe en finés, francés y ruso (véase Fig. 105). En dichas pruebas no sólo se utilizó la probabilidad condicional simétrica (la única empleada en las pruebas oficiales) sino también la información mutua; se experimentó con distintas medidas de similitud además de S_2 y se empleó el método de ponderación interdocumental basado en el coeficiente de variación que se desarrolló con posterioridad a *CLEF'04*.

A la vista de los resultados obtenidos parece que (1) la información mutua resulta sustancialmente mejor que la probabilidad condicional simétrica como método para calcular la significatividad de los *n*-gramas, (2) S_2 resulta sistemáticamente la mejor medida de similitud con independencia del idioma (si se emplea información mutua) y (3) el método de ponderación interdocumental permite, en general, mejorar sustancialmente los resultados. En cuanto al rendimiento de *blindLight* comparado con el de otros sistemas *IR*, estos datos son consistentes con los obtenidos en las pruebas oficiales.

En resumen, la utilización de *blindLight* como técnica de recuperación de información es viable ofreciendo, además, un método de pseudo-traducción de consultas que lo hace muy interesante para entornos multilingües. Ciertamente, los resultados obtenidos por el momento no son tan satisfactorios como los que proporcionan técnicas afianzadas; no obstante, son similares a los de nuevas técnicas consideradas "prometedoras" (p.ej. indexado por semántica latente) y se han señalado una serie de puntos donde, sin duda, será posible mejorar la técnica alcanzando rendimientos superiores.

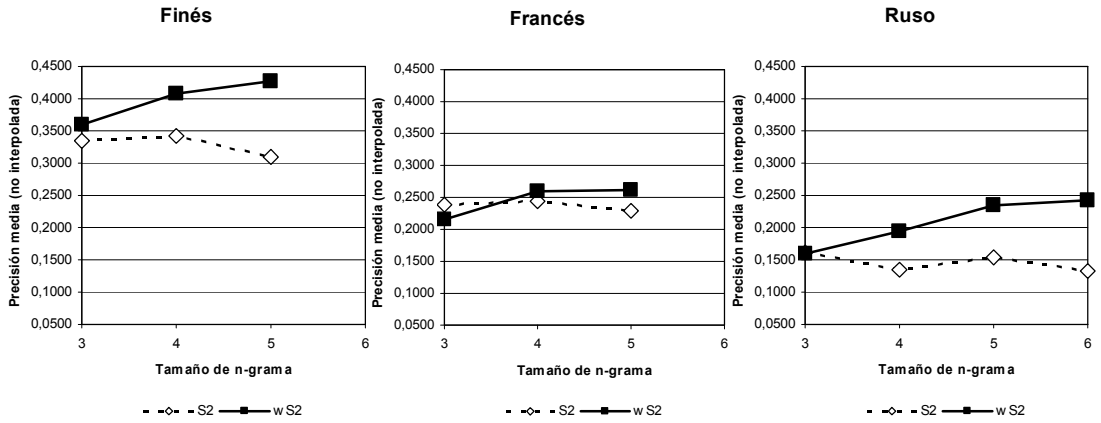


Fig. 106 Evolución de la precisión media (no interpolada) a medida que aumenta el tamaño de *n*-grama empleado para indexar las colecciones.

[1] Son pruebas oficiales aquellas en las que los participantes envían, para ser evaluadas por la organización del *CLEF*, listas de documentos que satisfacen el conjunto de consultas de la campaña en curso. Las pruebas oficiales se llevan a cabo sin que los equipos conozcan los juicios de relevancia. La diferencia entre pruebas oficiales y no oficiales radica tan sólo en el hecho de que en estas últimas es el investigador (y no la organización) quien calcula el rendimiento de su sistema a partir de los juicios de relevancia (mediante el programa `treceval`).

CAPÍTULO 7

EXTRACCIÓN DE RESÚMENES CON *BLINDLIGHT*

Las técnicas de resumen automático tienen como misión obtener a partir de un documento o conjunto de documentos más o menos estructurados un único texto mucho más corto que aún contenga los aspectos más relevantes de los originales. Este tipo de técnicas deben trabajar con distintos idiomas y su salida tendría que ser configurable por el usuario no sólo en cuanto a su longitud sino en función de sus necesidades de información. El resumen automático de textos resulta inestimable para aquellos usuarios que tratan con grandes cantidades de documentos y precisan de una herramienta que les permita determinar la información más relevante de un texto o un conjunto de textos a fin de discrimar aquellos a los que dedicar su atención. Durante los años 1950 y 1960 la investigación en este tipo de tecnologías fue intensa para descender considerablemente durante los años siguientes y no recuperarse hasta los años 1990. Desde entonces se trata de un campo muy activo y aunque aún se está lejos de disponer de sistemas capaces de emular a un ser humano (p.ej. produciendo resúmenes indicativos[1]) se ha avanzado enormemente y los sistemas estadísticos y puramente extractivos han demostrado su utilidad. En este capítulo se repasará brevemente la investigación desarrollada hasta el momento en este campo, prestando especial atención al enfoque estadístico. Se describirá la utilización de blindLight como sistema de resumen extractivo y se presentarán los resultados de la evaluación del mismo en un marco estandarizado, resultados que demuestran que la técnica propuesta por el autor resulta más eficaz que muchos de los métodos más avanzados disponibles.

1 Resumen automático

A lo largo de esta disertación se ha mencionado la sobrecarga de información que sufren los usuarios al tratar de localizar información textual y se han señalado distintas tareas que pueden paliar dicho problema como la categorización y clasificación de documentos, la recuperación de información o el resumen automático. Esta última ha atraído mucha

[1] Un **resumen indicativo** tan sólo sugiere el tema de un documento sin revelar sus contenidos.

atención durante los últimos años[1] dado su interés en un entorno con múltiples fuentes que, en mayor o menor medida, se solapan proporcionando mucha información redundante (véase Fig. 107).

Fig. 107 Noticias "agregadas" por *Google* (`http://news.google.com`).

El número de fuentes de las que un usuario puede obtener información es extraordinariamente elevado (existen 1.256 artículos relativos a la primera noticia y 430 para la segunda). Puesto que siempre existirá un "solapamiento" de contenidos sería necesario extraer la información más relevante de las distintas fuentes y proporcionarla de una forma integrada al usuario. Las técnicas de obtención de resúmenes a partir de múltiples documentos tienen un papel muy importante en este escenario.

El **resumen automático** puede definirse como el conjunto de técnicas que producen a partir de un texto de entrada un documento de salida de menor extensión pero que aún contiene los puntos más relevantes del original. Los orígenes de este campo de estudio pueden remontarse a los trabajos de Luhn (1958) y Edmundson (1969) que desarrollaron los primeros sistemas de "extracción de resúmenes".

Dichos sistemas (de **resumen extractivo**) construían los resúmenes a partir de sentencias extraídas del documento original en función de una serie de heurísticos como la utilización de palabras clave o que apareciesen en el título, la posición de las sentencias dentro del documento o la ausencia de palabras "estigma" (p.ej. conjunciones o pronombres al comienzo de la sentencia).

En el extremo opuesto se situarían sistemas de resumen automático que sintetizan un texto totalmente nuevo que recoge las ideas principales del documento original sin incluir necesariamente fragmentos literales del mismo, es decir, estos sistemas trabajarían de manera similar a un ser humano (**resumen abstractivo**). No obstante, el método "abstractivo", mucho más complejo, no es abierto puesto que requiere un importante conocimiento del dominio en que se realizan los resúmenes (Spärck-Jones 1999) y, de hecho, no es previsible la existencia de sistemas prácticos de resumen por abstracción a corto plazo (Hovy 1999, p. 7).

Así pues, la mejora de los métodos de resumen extractivo es un campo de investigación activo debido, por un lado, a las menores exigencias de partida (requieren un

[1] Durante la última década se han celebrado varios "talleres" y se han establecido conferencias dedicadas en exclusiva a las técnicas de resumen automático como *DUC – Document Understanding Conferences* (`http://duc.nist.gov/`).

conocimiento lingüístico nulo o mínimo) y, por otro, al hecho de que la mayor parte de documentos con los que tratan los usuarios en la actualidad no tienen una estructura fija ni pertenecen a un dominio concreto. En semejante escenario la sencillez, flexibilidad y robustez de los métodos de extracción son aspectos valiosos.

Luhn (1958) fue el primero en proponer un método estadístico para extraer las sentencias más significativas de un texto y construir un resumen del mismo. Luhn proponía determinar en primer lugar la significatividad de las distintas palabras[1], suponiendo que las más frecuentes (a excepción de las palabras vacías) serían las más importantes. Posteriormente se asignaría a cada sentencia un peso en función del número de palabras relevantes que incluyese, el peso de las mismas y la distancia entre ellas dentro de la sentencia. Una vez obtenida la puntuación de todas las sentencias del documento sería posible ordenarlas de mayor a menor importancia y seleccionar un subconjunto de las más significativas como resumen del texto original. Resulta interesante notar que Luhn también vio la necesidad de adaptar los resúmenes automáticos a los distintos intereses de los usuarios; proponía para ello asignar una "prima" a las palabras utilizadas por el usuario para describir su necesidad de información de tal modo que las sentencias que contuviesen dichas palabras obtuviesen mayores puntuaciones y pasasen al resumen con mayor facilidad.

Edmundson (1969) emplea cuatro métodos distintos para asignar pesos a las sentencias del documento: (1) un diccionario con palabras que proporcionan pistas sobre la relevancia (*bonus*) o irrelevancia (*estigma*) de las sentencias, (2) la utilización de palabras frecuentes (y no vacías) como indicadores de relevancia, (3) el uso de palabras del título del documento y de los apartados como indicadores positivos y (4) heurísticos basados en la posición de las sentencias en el texto[2]. Cada uno de estos métodos contribuía al peso final de las sentencias de manera independiente y configurable. Cabe señalar que Edmundson fue uno de los primeros en señalar la necesidad de evaluar los sistemas de extracción de resúmenes comparando sus resultados con resúmenes producidos por evaluadores humanos.

Durante los años 1970 y primeros 1980 la investigación en sistemas de resumen automático descendió considerablemente (Spärck-Jones 1999, p. 2) y no resurgiría hasta finales de los 1980 y en especial a partir de los 1990 siendo desde entonces un campo muy activo. Durante este tiempo se abordaron nuevos métodos más allá de la simple extracción de pasajes, por ejemplo, (DeJong 1982), (Tait 1983), (Fum, Guida y Tasso 1985) o (Reimer y Hahn 1988).

No obstante, muchos de estos sistemas tan sólo posibilitaban la generación de resúmenes mediante "plantillas" predefinidas que se rellenaban con datos extraídos de los documentos a resumir y que debían pertenecer a un número de géneros muy limitado (p.ej. Paice y Jones 1993). La técnica del autor, no obstante, sigue un enfoque extractivo puramente estadístico y, en consecuencia, a partir de este punto tan sólo se revisarán algunos de los principales trabajos desarrollados siguiendo métodos análogos; el lector interesado en el enfoque abstractivo puede acudir en primer lugar al capítulo cuarto del libro *"Advances in Automatic Text Summarization"* (Mani y Maybury, eds. 1999).

---

[1] El artículo de Luhn señala la necesidad de llevar a cabo *stemming* a fin de no diferenciar las distintas variantes de un mismo concepto.

[2] Por ejemplo, las sentencias que siguen a un título suelen ser relevantes y las sentencias relevantes tienden a aparecer muy pronto o muy tarde en el documento y en cada párrafo.

Salton y Singhal (1994) aplicaron algunas de las características del sistema de recuperación de información *SMART* a la obtención de resúmenes automáticos. Señalan la necesidad de identificar en primer lugar los distintos temas tratados en un documento así como los párrafos del texto que se refieren al mismo asunto para, posteriormente, emplear una selección de párrafos como resumen del documento. Básicamente se trata de realizar una clasificación automática de párrafos controlada mediante un umbral de similitud que será inversamente proporcional al número de temas que se deseen "descubrir" en el documento. Una vez clasificados los párrafos se extraen en tripletes de similitud máxima hasta alcanzar el tamaño de resumen deseado por el usuario colocándose en el mismo orden en que se encontraban en el texto original.

Salton, Singhal, Buckley y Mitra (1996) ahondan en el mismo tema pero tratando de avanzar desde la identificación (y clasificación) de párrafos hacia la identificación de **pasajes**, esto es, *"fragmentos de texto que exhiben consistencia interna y que pueden distinguirse del resto de texto circundante"* (Salton *et al.* 1996, p. 3). Esta iniciativa es similar a la de la técnica de *TextTiling* (Hearst 1994) con la salvedad de que Hearst no explicitó el interés de la misma para la extracción automática de resúmenes. Salton, Singhal, Mitra y Buckley (1997) desarrollan aún más algunas de las ideas expuestas por Salton y Singhal (1994) y Salton *et al.* (1996) sobre la utilización del grafo de relaciones entre pasajes para la extracción de aquellos más significativos y la construcción de un resumen automático.

Mitra, Singhal y Buckley (1997) evalúan el anterior sistema comparándolo con resúmenes extractivos creados manualmente. En su estudio llegan a una serie de conclusiones que todavía son vigentes: (1) los primeros párrafos de un texto resultan tan efectivos como un resumen obtenido mediante métodos extractivos "inteligentes", (2) esto puede deberse a que los documentos normalmente empleados en los experimentos (artículos periodísticos, técnicos o científicos) están estructurados de tal modo que los primeros párrafos ya son un resumen lo cual explicaría los buenos resultados del *baseline*[1] y (3) aunque *"el resumen por extracción es un método imperfecto parece ser la única técnica que funciona razonablemente con independencia del dominio"*. Brandow, Mitze y Rau (1995) desarrollaron un trabajo similar llegando a conclusiones parecidas, en particular, que los usuarios preferían los resúmenes producidos por el método *baseline*; no obstante, Zechner (1996) señaló que, aunque tal vez menos legibles, los resultados de las técnicas automáticas son mejores en términos de precisión y exhaustividad.

El trabajo de Kupiec, Pedersen y Chen (1995) resulta muy interesante puesto que emplearon un *corpus* de documentos y resúmenes creados manualmente como datos de entrenamiento para un clasificador bayesiano que debía determinar qué sentencias de un documento deberían formar parte de un resumen y cuáles no. El sistema propuesto determinaba para cada sentencia la probabilidad de pertenencia al resumen final y extraía las más probables. Al reducir los documentos a un 25% del tamaño original seleccionaba un 84% de las sentencias elegidas por los expertos humanos y para resúmenes más cortos resultaba sustancialmente superior al *baseline* consistente en presentar el inicio del documento.

Más recientemente, Kraaij, Spitters y van der Heijden (2001) y Kraaij, Spitters y Hulth (2002) también han utilizado clasificadores bayesianos para la extracción de resúmenes. Por su parte, Conroy *et al.* (2001) y Dunlavy *et al.* (2003) han implementado sistemas extractivos mediante modelos de Markov que hacen menos suposiciones que los

[1] Un método *baseline*, literalmente "línea base", es una técnica trivial para resolver un problema en estudio y frente a la cual se comparan los resultados obtenidos mediante las nuevas propuestas.

clasificadores bayesianos sobre la independencia entre elementos. Hirao *et al.* (2002 y 2003) han empleado *SVM's* de manera similar con bastante éxito y Fuentes *et al.* (2003) o Doran *et al.* (2004) árboles de decisión. Alfonseca y Rodriguez (2003), Jaoua y Ben Hamadou (2003, citado por Alfonseca *et al.* 2004) y Alfonseca, Guirao y Moreno Sandoval (2004) han utilizado algoritmos genéticos para la selección de las sentencias.

Fukumoto, Suzuki y Fukumoto (1997) asignan a cada palabra del texto un peso que dependerá de su distribución en el propio documento y en un contexto más amplio. Según su criterio una palabra será palabra clave si (1) su dispersión a nivel de párrafo es menor que a nivel de documento y (2) ésta a su vez es menor que la del término en el dominio. Estos criterios se implementan utilizando el método de ponderación χ^2 de Watanabe *et al.* (1996). Así, para cada palabra no vacía se determina su peso χ^2 dentro del párrafo, el documento y el dominio y se seleccionan aquellas que verifican los dos criterios anteriormente expuestos. Posteriormente cada párrafo del documento se representa mediante un vector que sólo incluirá las correspondientes palabras clave y se realiza una clasificación automática de manera análoga a la de Salton *et al.* (1994, 1996 y 1997). El resumen se construirá seleccionando en primer lugar aquellos párrafos que estén incluidos en un mayor número de los grupos resultantes del proceso de clasificación. La principal ventaja de este método radica en la posibilidad de ajustar los resúmenes a distintos contextos pero, al mismo tiempo, es su principal inconveniente al requerir un *corpus* para extraer resúmenes.

Hovy y Lin (1997) describen el sistema *SUMMARIST* que trata de integrar los enfoques extractivos y abstractivos mediante un proceso de tres fases: (1) identificación de tópicos, (2) interpretación y (3) generación. La primera fase se basa en la denominada Política de Posición Óptima (*Optimal Position Policy*) que no es más que una lista que señala las posiciones donde es más probable encontrar los aspectos clave de un texto de un género determinado[1]. Aplicando únicamente esta primera fase de identificación sería posible construir resúmenes extractivos; no obstante, Hovy y Lin plantean las fases de interpretación y generación para evitar algunos de los problemas de este tipo de resúmenes[2]. Sugieren, por ejemplo, emplear *WordNet*[3] en la segunda fase de tal modo que se puedan resolver situaciones como la que se muestra en Fig. 108.

John bought some **vegetables, fruit, bread and milk.**
John bougth some **groceries.**

Fig. 108 Ejemplo de situación que se resolvería en la fase de "interpretación" (Hovy y Lin 1997).

En (Lin y Hovy 2000) se describe otro concepto interesante para las fases de interpretación y generación: las denominadas *topic signatures*. Estas no son más que conjuntos de términos relacionados, susceptibles de ser reemplazados en el resumen final por un único concepto y obtenibles por medios puramente estadísticos a partir de un *corpus*. Por ejemplo, si los términos `mesa`, `menú`, `camarero`, `comida`, `propina`, etc. apareciesen combinados en un documento podrían sustituirse por la frase `visita a restaurante` en el momento de construir el resumen. *NeATS* (Lin y Hovy 2001 y 2002a) es un sistema de resúmenes

[1] Por ejemplo, según Hovy y Lin la política para el *Wall Street Journal* sería `[T1, P1S1, P1S2, …]`, o lo que es lo mismo, los aspectos más relevantes del documento aparecen en el título, la primera sentencia del primer párrafo seguido de la segunda sentencia del primer párrafo, etc. Otros dominios tendrían políticas distintas que habría que descubrir. Lin y Hovy (1997) describe en detalle el modo en que es posible obtener de modo automático una de tales políticas.

[2] Lal y Ruger (2002) han realizado un trabajo similar en lo referente a la simplificación de las sentencias extraídas.

[3] `http://wordnet.princeton.edu`

multidocumento que aplica las ideas anteriores y que obtuvo interesantes resultados en las campañas *DUC* (*Document Understanding Conferences*) de 2001 y 2002.

Barzilay y Elhadad (1997) plantean la utilidad de las cadenas léxicas como elemento facilitador en la extracción de resúmenes. Una **cadena léxica** es una secuencia de palabras semánticamente relacionadas que aparecen en un texto y que pueden ser adyacentes o encontrarse dispersas a lo largo del documento. Para encontrar dichas cadenas léxicas en un texto genérico es necesario utilizar recursos como *WordNet* que proporcionan la información necesaria sobre las posibles relaciones entre distintas palabras. Así pues, Barzilay y Elhadad encuentran en primer lugar cadenas léxicas en el texto, seguidamente asignan a cada cadena léxica una puntuación[1] y, por último, seleccionan aquellas sentencias que mejor satisfacen a las cadenas léxicas de mayor puntuación. Brunn, Chali y Pinchak (2001) desarrollaron un trabajo muy similar concluyendo también que las cadenas léxicas pueden resultar muy interesantes para introducir conocimiento lingüístico en los métodos extractivos. McKeown *et al.* (2001), Fuentes *et al.* (2003) y Doran *et al.* (2004) también han utilizado cadenas léxicas como método de puntuación.

Jing y McKeown (2000) de la Universidad de Columbia estudiaron diversas tareas de post-procesamiento de los resúmenes extractivos para mejorar su calidad. Aun cuando otros autores (Mani, Gates y Bloerdon 1999) ya trataron dicho problema el interés de este trabajo radica en la forma en que se aborda: Jing y McKeown desarrollaron una técnica que permite en primer lugar analizar la relación entre un resumen manual (creado por un humano) y el documento original a fin de determinar, por un lado, las sentencias "extraídas" y, por otro, las fases de reducción, combinación y reordenamiento a que fueron sometidas. De este modo, no sólo obtienen un conocimiento muy interesante sobre la forma en que un ser humano crea resúmenes mediante "corta-y-pega", sino que son capaces de entrenar su sistema a fin de que emule, hasta cierto punto, estas capacidades. Su equipo ha obtenido muy buenos resultados en *DUC* con el método extractivo (McKeown *et al.* 2001 y 2002) aunque en las últimas ediciones tiende más hacia el enfoque generativo al fusionar y reescribir las sentencias extraídas (Nenkova *et al.* 2003). No obstante, para la obtención de resúmenes a partir de texto traducido automáticamente siguen optando por la utilización de técnicas extractivas (Blair-Goldensohn *et al.* 2004).

Hardy *et al.* (2001) describen un sistema para construir resúmenes a partir de varios documentos (decenas o cientos). Para ello, dividen cada documento en párrafos que son clasificados automáticamente empleando una medida de similitud basada en *n*-gramas de palabras. Una vez descubiertos los distintos grupos se selecciona un párrafo de cada uno para construir el documento final.

MEAD (Radev, Blair-Goldensohn y Zhang 2001) también es un sistema extractivo para la obtención de resúmenes a partir de múltiples documentos. Para cada sentencia de cada documento del conjunto a resumir el sistema obtiene tres puntuaciones a partir de (1) la similitud entre la sentencia y el centroide del conjunto, (2) la distancia de la sentencia al inicio de su correspondiente documento y (3) la similitud entre la sentencia y la primera sentencia (o el título) del documento al que pertenece. Estas puntuaciones se normalizan en el intervalo [0, 1] y se combinan linealmente para obtener una única puntuación que permita

[1] Barzilay y Elhadad (1997) determinaron empíricamente que son dos los parámetros de una cadena léxica que resultan buenos predictores acerca de su utilidad para la construcción de un resumen: la "longitud" y el "índice de homogeneidad" entendidas, respectivamente, como el número de ocurrencias en el texto de miembros de la cadena y 1 - el número de distintas ocurrencias dividido entre la longitud. Así, la puntuación de una cadena léxica sería el producto de su longitud por su índice de homogeneidad.

seleccionar las sentencias más relevantes. Por último, se eliminan del resumen aquellas sentencias demasiado similares entre sí. Posteriormente fue adaptado para la extracción de resúmenes guiados por preguntas resultando el mejor participante en dicha tarea de *DUC 2003* (Radev *et al.* 2003). Saggion y Gaizauskas (2004) han llevado a cabo un trabajo similar.

Recientemente, Erkan y Radev (2004a) han desarrollado una nueva medida de "centralidad" para las sentencias, denominada *LexPageRank*, basada en la idea de "prestigio" de las redes sociales y análoga al *PageRank* (Page *et al.* 1998) de *Google*. El valor de *LexPageRank* para una sentencia *S* se define como la suma de los valores *LexPageRank* de aquellas sentencias similares a *S*, donde la similitud se determina mediante la función del coseno. Esta última versión de *MEAD* resultó uno de los mejores participantes en cuatro de las cinco tareas de *DUC 2004* (Erkan y Radev 2004b). Por su parte, Vanderwende, Banko y Menezes (2004) utilizan *PageRank* para determinar qué elementos de un documento son los más relevantes aunque sus resúmenes son generados y no construidos a partir de sentencias extraídas literalmente de los documentos. Ambos trabajos guardan cierta relación con los desarrollados por Salton *et al.* (1996) que también emplearon grafos para analizar los contenidos de un texto.

En resumen, aunque la calidad de los resultados de los métodos puramente extractivos puede ser deficiente en ocasiones, lo cierto es que estas técnicas son mucho más flexibles y generales que las abstractivas (Spärck-Jones 1999) y existen toda una serie de métodos de post-procesamiento sencillos y capaces de mejorar enormemente la legibilidad del texto final.

2 Utilización de *blindLight* para la extracción de resúmenes[1]

Como se recordará, *blindLight* es una técnica bioinspirada (véase pág. 59) construida sobre la idea de un "genoma documental" definido así:

> *El ADN de un documento es un conjunto de genes donde cada gen está formado por un n-grama de caracteres y su correspondiente significatividad dentro del documento de origen.*

En los capítulos anteriores se ha visto cómo se pueden comparar los genomas de distintos documentos (o consultas) para desarrollar sistemas de clasificación (pág. 85), categorización (pág. 117) y recuperación de información (pág. 141). Estas comparaciones requieren la previa "intersección" de dos pares de "secuencias génicas" y se ha mostrado cómo esta misma operación puede resultar útil para desarrollar una pseudo-traducción de textos (pág. 142). Sin embargo, esta idea de un "genoma documental" aún puede llevarse un paso más allá.

En la naturaleza el ADN codifica los distintos aminoácidos que son los elementos constituyentes básicos de las proteínas, las cuales, a su vez, interpretan un papel esencial en la práctica totalidad de procesos biológicos. Así, las células producen las proteínas necesarias en cada momento empleando el ADN a modo de "plano de construcción". No obstante, el ADN no es muy versatil químicamente y, más aún, es demasiado valioso como para trabajar directamente sobre su molécula para construir cada proteína. Por esa razón, las porciones de ADN con el gen o genes que codifican cada proteína son previamente copiadas mediante moléculas de ARN mensajero (ARNm). El ARNm sale del núcleo celular hacia el citoplasma donde los ribosomas lo emplean como plantilla sobre la que se construye, aminoácido a aminoácido, la proteína utilizando ARN transferente (ARNt). El proceso

[1] La técnica aquí descrita es una evolución de la presentada por Gayo Avello *et al.* (2004a).

durante el cual se copia una porción de ADN sobre ARNm se denomina transcripción y la fase en que se construye la cadena de aminoácidos a partir de la molécula de ARNm se conoce como traducción (véase Fig. 109).

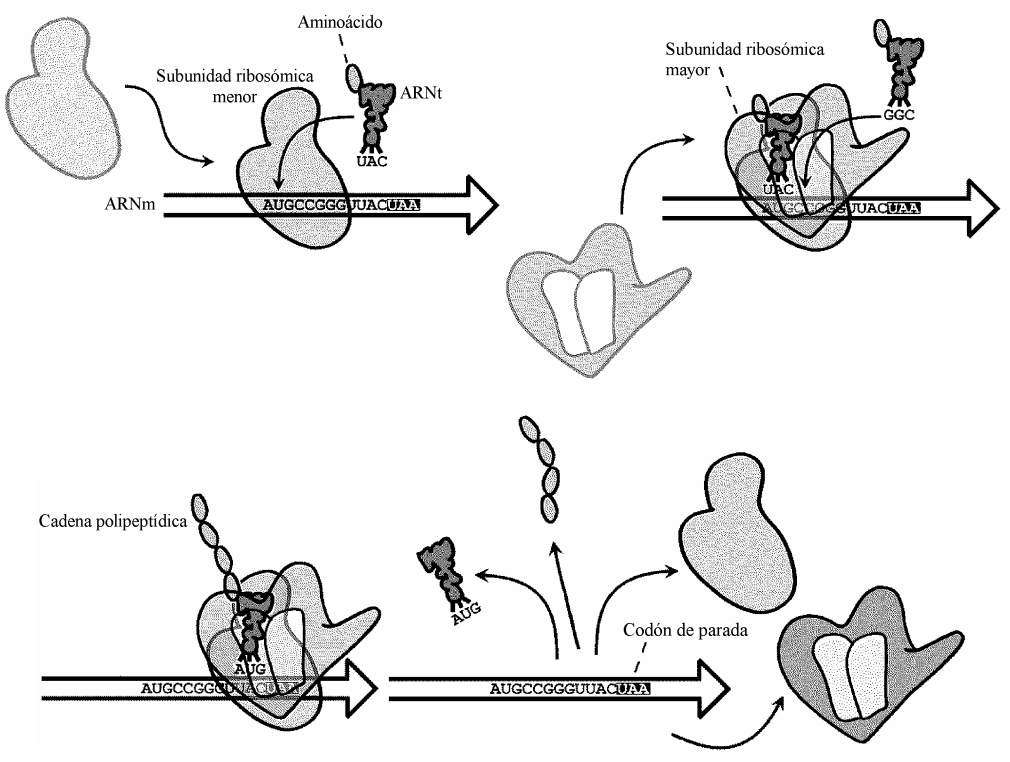


Fig. 109 (De izquierda a derecha y de arriba abajo) **Inicio de la traducción, comienzo de la elongación, fin de la elongación, fin de la traducción.**

La síntesis de una proteína comienza con la unión de la subunidad ribosómica menor a la cadena de ARNm. Entonces la molécula de ARNt iniciador se une al codón de inicio AUG. La subunidad ribosómica mayor es atraída por el ARNt iniciador completando el ribosoma que comenzará a desplazarse a lo largo del ARNm codón a codón. Cada uno de los codones en el ARNm tiene un anticodón complementario en las correspondientes moléculas de ARNt. Cada una de estas moléculas transporta un aminoácido que es añadido a la creciente cadena polipeptídica. Al llegar al codón de parada finaliza el proceso de traducción, el ribosoma se disgrega y la cadena polipeptídica queda libre, plegándose y formando la proteína final.

La utilización de *blindLight* como técnica de extracción de resúmenes se inspira en este proceso de traducción y síntesis de las proteínas para lo cual emplea tanto el vector de *n*-gramas obtenido a partir del documento como el texto plano original del mismo. La ideas subyacentes son muy sencillas:

1. El "ADN documental" está codificado mediante un vector de *n*-gramas de caracteres, cada uno de los cuales tiene asociado un peso, su significatividad. Cada uno de estos pares `(n-grama, significatividad)` puede emplearse a modo de ARNt (véase Fig. 110).
2. El texto plano no proporciona ninguna información al ordenador sobre la relevancia de los distintos pasajes. Sin embargo, puede procesarse de manera secuencial y junto con el "ARNt documental" se puede transferir significatividad a dicho texto (véase Fig. 111).
3. El proceso de transferencia de significatividad del "ARNt documental" al texto no se realiza en una única fase sobre el texto completo sino en varias pasadas

garantizando que la significatividad media por carácter sea creciente. De este modo, el texto de partida es "troceado" en fragmentos (*chunks*) de máxima significatividad. Dichos fragmentos pueden ser utilizados posteriormente para obtener palabras clave o para facilitar la extracción de las sentencias más relevantes (véase Fig. 112 y Fig. 113).

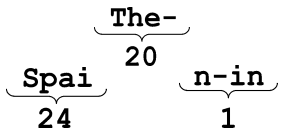


Fig. 110 Cada componente del vector de *n*-gramas puede utilizarse a modo de ARNt.

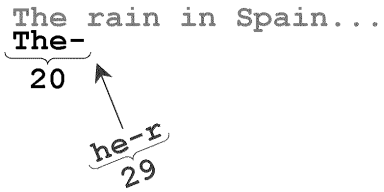


Fig. 111 El "ARNt documental" permite asociar al texto información sobre su significatividad.

El proceso en sí mismo también es muy simple. En primer lugar debe obtenerse un vector de *n*-gramas con sus correspondientes significatividades tal y como se hace para cualquier otra aplicación de la técnica. Una vez hecho esto se pasa a trabajar sobre el texto original que se habrá segmentado en frases y sentencias[1]. Tal segmentación es necesaria por dos razones: (1) al dividir el texto en fragmentos de máxima significatividad no se debe saltar entre dos frases y (2) es necesario conocer los límites entre sentencias para garantizar una mínima coherencia en el resumen extraído.

La Comisión ha adoptado hoy propuestas relativas a un paquete de medidas destinadas a reforzar la capacidad de respuesta de la Unión Europea en caso de catástrofes. Estas medidas se destinan a financiar nuevos equipos especializados en materia de planificación para agilizar el suministro eficaz de ayuda a largo plazo; a reforzar la capacidad de la Unión de facilitar equipos de expertos civiles y de equipo y a suministrar ayuda humanitaria. La Comunicación adoptada hoy también presenta un informe detallado sobre la utilización de los 450 millones de euros anunciados por la UE tras la catástrofe del tsunami. Las propuestas adoptadas hoy constituyen la contribución de la Comisión al plan de acción tras el tsunami propuesto por la Presidencia luxemburguesa el 31 de enero.

«Vistas las situaciones anteriores y nuestra capacidad de responder inmediatamente ante la catástrofe del tsunami, la Comisión propone ahora medidas que nos ayudarán, en el futuro, a contribuir de forma rápida y eficaz a las tareas de reconstrucción tras una catástrofe» ha declarado la Comisaria de Relaciones Exteriores y Política de Vecindad, Benita Ferrero-Waldner, que propone dichas medidas conjuntamente con los Comisarios Michel y Dimas. Stavros Dimas, Comisario Europeo responsable de Protección Civil ha dicho: «Nuestra reacción ante el Tsunami ha demostrado el claro valor añadido que la dimensión europea aporta a la asistencia en materia de protección civil. Las propuestas de hoy hacen avanzar un paso más al Mecanismo actual... Tomadas en su conjunto, permitirán disponer de un instrumento que garantiza una reacción europea eficaz ante futuras catástrofes».

Fig. 112 División de un texto en fragmentos de máxima significatividad.

Con fondo gris se muestran los 50 fragmentos más significativos, con borde negro los 10 más significativos.

Una de las primeras tareas a realizar sobre el texto del documento es la obtención de la significatividad media por carácter de cada sentencia que no es más que el cociente de la suma de las significatividades de todos los *n*-gramas que aparecen en la sentencia entre la longitud de la misma. Este valor será una de las varias "puntuaciones" que se utilizarán para determinar la relevancia de las distintas sentencias.

La siguiente fase es la ya citada división del texto en fragmentos de máxima significatividad, fragmentos que, como también se ha dicho, deben estar enteramente contenidos en una única frase. El algoritmo para llevar a cabo esta fragmentación se muestra

[1] El modo en que se lleve a cabo la segmentación es irrelevante. Para utilizar *blindLight* tan sólo se precisa de sentencias y frases entendiendo las primeras como oraciones gramaticales y las segundas como secuencias de palabras dotadas de sentido pero que no forman oración.

en Fig. 115; no obstante se describirá brevemente a continuación y se proporciona un ejemplo ilustrativo en Fig. 116.

os 450 m	la Comisión p	catástrofes
luxemburgues	e la Comisión	catástrofe
trofe» h	agilizar	a Comisión
Políti	el 31 de	la Comisión
también p	La Comisión h	el tsunam
catástrofes»	ión Europe	eficaz
ro-Wald	el Tsunam	medidas
catástrofes	tilización	equipos
hoy propuest	conjunt	la capacida
catástrofe	nstrucción	cción

Fig. 113 (A la izquierda) **20 fragmentos más significativos de un documento y** (a la derecha) **10 primeras "palabras clave" obtenidas al desplazar una ventana sobre los primeros.**

Este algoritmo trabaja sobre dos estructuras diferentes: por una parte, una lista que inicialmente contiene las frases extraídas del texto original y, por otra, una pila con los *n*-gramas del documento ordenados por significatividad decreciente. En cada iteración se extrae un *n*-grama de la pila, que será el *n*-grama más significativo del texto disponible en la lista de frases, y se buscan aquellas frases que lo contengan. Posteriormente, para cada una de las frases se localizan los *n*-gramas anterior y posterior al fragmento localizado y se aumenta dicho fragmento añadiéndole el *n*-grama más significativo del par. Este proceso se repite para cada frase extraída mientras la significatividad por carácter no decrezca. En el momento en que el fragmento de texto no pueda crecer sin disminuir su significatividad se detiene la fase de crecimiento, se extrae el fragmento y se elimina la frase de la lista, sustituyéndola por las secciones anterior y posterior al fragmento extraído. Una vez se ha terminado de procesar todas las frases correspondientes a un *n*-grama se repite todo el proceso para el siguiente *n*-grama finalizando cuando la pila quede vacía. En ese momento se habrá segmentado todo el texto del documento en fragmentos de máxima significatividad.

Una vez hecho esto se puede obtener una nueva puntuación para cada sentencia; dicha puntuación no es más que la suma de la significatividad media por carácter de cada fragmento presente en la sentencia. A partir de la lista de fragmentos también es posible obtener las "palabras clave" del documento. Para ello basta con recorrerla con una ventana de tamaño *K* extrayendo como claves las subcadenas más largas que resulten de la intersección de cualquier par de fragmentos contenidos en la ventana (véase Fig. 114).

la Comisión	catástrofes	catástrofes	catástrofes
el tsunam	catástrofe	catástrofe	catástrofe
medidas	a Comisión	a Comisión	a Comisión
cción	la Comisión	la Comisión	la Comisión
uest	el tsunam	propuest	la Comis
teri	eficaz	el tsunam	a Comis
ión	medidas	eficaz	eficaz a
uro	equipos	medidas	propuest
ida	cción	equipos	el tsunam
aci	suministr	acción	La Com
	trofe	la capacida	civil
	ción	cción	eficaz
	Com	iliza	Comisa
	la U	suministr	Comis
	uest	ación	medidas

Fig. 114 (De izquierda a derecha) **15 primeras "palabras clave" obtenidas con ventanas de tamaño 2, 4, 8 y la longitud total de la lista de fragmentos. El peso de cada palabra clave es su significatividad por carácter.**

Naturalmente, las claves obtenidas pueden ser fragmentos de palabras o frases (véase Fig. 113), no obstante, esta solución es razonablemente flexible puesto que en el caso de textos pertenecientes a idiomas occidentales siempre se pueden refinar las claves encontradas (esto es, asegurarse de que se trata de palabras o frases completas) y para los idiomas orientales que no separan las palabras no requeriría una fase de segmentación previa. Sin embargo, la utilización de *blindLight* como sistema de extracción de palabras clave aún requiere una experimentación más rigurosa puesto que sus resultados aún no son todo lo satisfactorios que se desearía como se verá en próximos apartados.

Por último, del mismo modo que se puede asignar a cada sentencia una puntuación en función de los fragmentos que contiene también es posible elaborar otra puntuación partiendo de las palabras clave: este valor sería la suma de la significatividad media por carácter de cada palabra clave contenida en la sentencia.

Algoritmo ribosomalTranslation *(sentences, ngramStack, sizeNgram)*
Input: *sentences*, el texto del documento segmentado en sentencias, *ngramStack* una pila con los *n*-gramas del documento donde el tope contendrá el n-grama más significativo aún sin tratar y *sizeNgram*, el tamaño de *n*-grama empleado.

```
1.  while ngramStack ≠ λ do
2.      ngram ← pop (ngramStack)
3.      w ← ngramWeight (ngram)
4.      sentenceSubset ← searchSentencesContaining (sentences, ngram)
5.      for each sentence in sentenceSubset do
6.          chunk ← ngram
7.          totalSig ← w
8.          currentAvgSig ← 0
9.          do
10.             previousAvgSig ← currentAvgSig
11.             lead ← leadNgram (chunk, sentence, sizeNgram)
12.             tail ← tailNgram (chunk, sentence, sizeNgram)
13.             w_L ← ngramWeight (lead)
14.             w_T ← ngramWeight (tail)
15.             if (w_L > w_T)
16.                 chunk ← charAt (lead, 0) + chunk
17.                 totalSig ← totalSig + w_L
18.             else
19.                 chunk ← chunk + charAt (tail, sizeNgram - 1)
20.                 totalSig ← totalSig + w_T
21.             end if
22.             currentAvgSig ← totalSig / strlen (chunk)
23.         while (currentAvgSig > previousAvgSig)
24.         chunks (chunk) ← currentAvgSig
25.         sentenceFragments ← explode (sentence, chunk)
26.         for each newSentence in sentenceFragments do
27.             insertInto (sentences, newSentence)
28.         loop
29.     loop
30. loop
31. return chunks
```

Fig. 115 Algoritmo que divide el texto del documento en fragmentos de máxima significatividad.

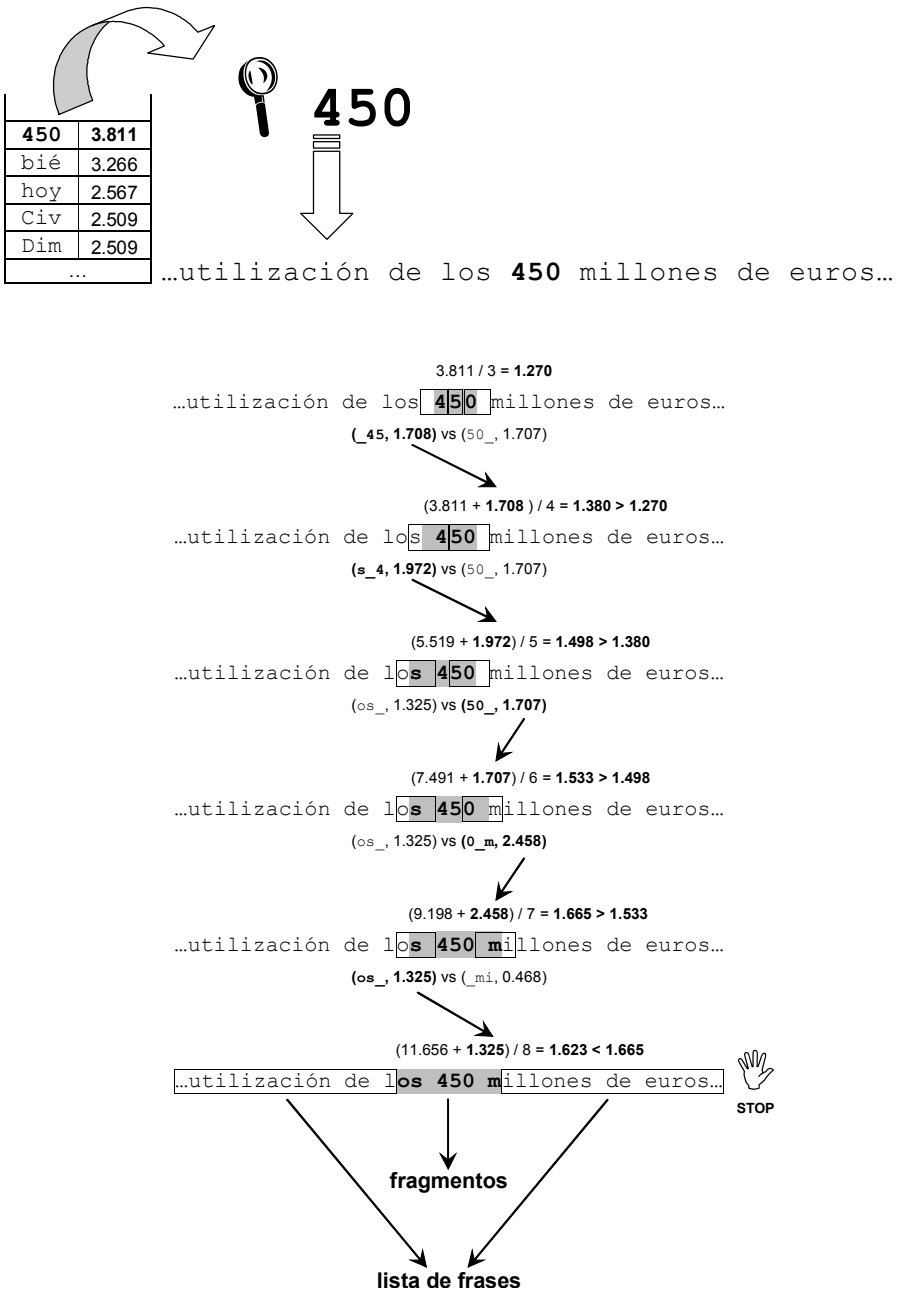


Fig. 116 Proceso mediante el cual se divide el texto en fragmentos de máxima significatividad.

Así pues, a partir del vector de *n*-gramas y del texto del documento segmentado en frases y sentencias es posible (1) obtener la significatividad media por carácter de cada sentencia, (2) segmentar el texto original en fragmentos de máxima significatividad, (3) utilizar dichos fragmentos para obtener una segunda puntuación para cada sentencia, (4) extraer a partir de los fragmentos las palabras clave del documento y (5) emplear las palabras clave para calcular una tercera puntuación para cada sentencia del texto. De acuerdo con el enfoque extractivo "clásico" (Edmundson 1969) para construir el resumen del documento tan sólo hay que extraer las sentencias con mayor puntuación mientras no se supere el tamaño máximo del resumen y colocarlas en el mismo orden en que aparecen en el texto original (véase Fig. 117). Por el momento la utilización de *blindLight* como sistema de extracción de resúmenes queda limitado a su utilización con un único documento; aún no se

ha implementado ningún sistema de resumen guiado mediante consultas y tan sólo se ha desarrollado una versión preliminar de extracción de resúmenes a partir de varios documentos que se describirá en un apartado posterior.

Esta técnica tiene un parecido superficial con la propuesta por Cohen (1995), *Highlights* (véase pág. 55). No obstante, la técnica presentada por el autor es, por un lado, más sencilla al no requerir el uso de un contexto externo y, por otro, más potente puesto que *Highlights* tan sólo permitía determinar (de una manera un tanto complicada) qué secuencias de caracteres eran relevantes para, posteriormente, extraer unos pocos términos clave mientras que la nueva técnica aquí descrita no sólo permite extraer dichos términos sino generar resúmenes extractivos de manera sencilla.

Significatividad media por carácter

La Comunicación adoptada hoy también presenta un informe detallado sobre la utilización de los 450 millones de euros anunciados por la UE tras la catástrofe del tsunami. Las propuestas de hoy hacen avanzar un paso más al Mecanismo actual... Tomadas en su conjunto, permitirán disponer de un instrumento que garantiza una reacción europea eficaz ante futuras catástrofes.

Puntuación por *chunks*

«Vistas las situaciones anteriores y nuestra capacidad de responder inmediatamente ante la catástrofe del tsunami, la Comisión propone ahora medidas que nos ayudarán, en el futuro, a contribuir de forma rápida y eficaz a las tareas de reconstrucción tras una catástrofe» ha declarado la Comisaria de Relaciones Exteriores y Política de Vecindad, Benita Ferrero-Waldner, que propone dichas medidas conjuntamente con los Comisarios Michel y Dimas.

Puntuación por palabras clave (ventana de tamaño 6)

La Comisión ha adoptado hoy propuestas relativas a un paquete de medidas destinadas a reforzar la capacidad de respuesta de la Unión Europea en caso de catástrofes. Las propuestas adoptadas hoy constituyen la contribución de la Comisión al plan de acción tras el tsunami propuesto por la Presidencia luxemburguesa el 31 de enero.

Significatividad media por carácter y puntuación por *chunks*

La Comisión ha adoptado hoy propuestas relativas a un paquete de medidas destinadas a reforzar la capacidad de respuesta de la Unión Europea en caso de catástrofes. La Comunicación adoptada hoy también presenta un informe detallado sobre la utilización de los 450 millones de euros anunciados por la UE tras la catástrofe del tsunami.

Fig. 117 Resúmenes del 20% (aprox.) empleando las distintas formas de puntuación descritas.

3 Evaluación de los sistemas de resumen automático

Medir de un modo objetivo la idoneidad de un resumen (automático o no) no es una tarea trivial. Según Hovy (1999) habría que determinar, al menos, dos medidas: la ratio de compresión (la relación entre la longitud del resumen y del texto original) y la ratio de retención[1] (cuánta información del original ha sido retenida por el resumen). Está claro que un buen resumen tendría ratios de compresión y retención próximas a cero y a uno respectivamente (Hovy y Lin 1998).

No obstante, mientras que calcular la primera es sencillo hacer lo propio para la segunda resulta mucho más complicado. Hovy y Lin (1998) describen tres tipos de experimentos que permitirían obtener sendas medidas de la ratio de retención. Se trata de los denominados juegos de Shannon, de la Pregunta y de Clasificación (*Shannon Game*, *Question Game* y *Classification Game*). En todos ellos era necesario recurrir a sujetos humanos que debían llevar a cabo una tarea que requería el conocimiento previo del texto original. Por ejemplo, en el caso del juego de Shannon los sujetos debían reconstruir el documento original de manera literal; algunos de los participantes habían tenido acceso al mismo mientras que otros sólo habían leído el resumen. En todos los casos se informaba a los sujetos cuando se equivocaban en una letra y se les permitía un nuevo intento; la relación

[1] En realidad, Hovy (1999) la denomina, paradójicamente, ratio de omisión pero con idéntico sentido.

entre el número de intentos requeridos en ambos grupos permitía calcular la ratio de retención.

No parece necesario decir que este tipo de experimentos son enormemente costosos en tiempo y recursos y, por otro lado, evalúan la "calidad" de los resúmenes de manera indirecta a través de su influencia en la ejecución de una o más tareas. Esto es lo que se conoce como evaluación extrínseca. No obstante, un sistema de resumen automático, y en general cualquier sistema PLN, puede ser evaluado también de manera intrínseca (Galliers y Spärck-Jones 1993). En este caso se trata de evaluar directamente la calidad del resúmen comparándolo con un resumen "modelo" pre-existente o creado a tal efecto por seres humanos. Este método ya fue reivindicado por Edmundson (1969) y es el más utilizado (Hovy 1999), por ejemplo, (Kupiec, Pedersen y Chen 1995), (Brandow, Mitze y Rau 1995), (Mitra, Singhal y Buckley 1997) o (Jing *et al.* 1998). A su vez, la evaluación intrínseca puede ser manual cuando la comparación entre resultados y modelo la realiza un ser humano o automática cuando se reemplaza al evaluador por algún tipo de algoritmo.

No obstante, no es suficiente con evaluar un sistema de resumen automático, es necesario que la evaluación sea aplicable a distintos sistemas y, en consecuencia, permita la comparación de diferentes técnicas. La serie de conferencias *DUC* (*Document Understanding Conferences*) surgió como una iniciativa[1] para el desarrollo de un marco común para la evaluación (y consecuente mejora) de los sistemas de resumen automático. Desde hace ya algún tiempo se trata del principal foro de evaluación de este tipo de sistemas, los documentos y modelos de ediciones anteriores están disponibles y desde 2004 la evaluación se hace de modo automático[2].

En todas las ediciones *DUC* se presentan diversas tareas que incluyen, entre otras, la obtención de resúmenes genéricos a partir de un único documento o de conjuntos de textos relativos a un tema común. La organización prepara los conjuntos de documentos y elabora los correspondientes resúmenes modelo para la posterior evaluación de los resultados.

Durante las ediciones 2001-2003 la evaluación se realizó de manera manual; es decir, una serie de jueces humanos debían "comparar" los resultados de los distintos participantes con los modelos disponibles. Por supuesto la comparación no era totalmente subjetiva sino que se realizaba con la asistencia de una herramienta (*SEE*)[3]. Inicialmente los textos a comparar se dividen en "unidades de discurso" (p.ej. frases) de tal manera que el revisor puede seleccionar distintas unidades del resumen a evaluar, asociarlas a unidades del modelo e indicar si los contenidos de la unidad en el resumen conciden total o parcialmente con aquellos de la unidad en el modelo. El revisor puede también indicar la calidad gramatical de cada unidad y, por último, evaluar de manera global la coherencia, gramática y organización del resumen automático. Finalmente, la herramienta calcula la coincidencia entre el resumen y el modelo como valores de precisión y exhaustividad (Lin y Hovy 2002b).

---

[1] Anteriormente ya se había realizado una experiencia de evaluación a gran escala, *SUMMAC* (Mani *et al.* 1998). Sin embargo, el enfoque seguido fue fundamentalmente extrínseco ya que se consideró que *"los resúmenes ideales son difíciles de conseguir y raramente únicos*". Por ello, el mérito de esta iniciativa no fue tanto la posibilidad de reutilizar sus productos para posteriores evaluaciones sino la demostración, por un lado, del interés de un marco de evaluación común e independiente de los desarrolladores y, por otro, de la enorme utilidad de las técnicas de resumen automático para otras tareas de tratamiento de información.

[2] Estos motivos han llevado al autor a evaluar su técnica mediante los datos de la última edición. Los resultados de dicha evaluación se presentan en un apartado posterior.

[3] *SEE* (*Summary Evaluation Environment*) disponible en: `http://www.isi.edu/~cyl/SEE/`

Harman y Over (2004) analizan los efectos de los distintos "factores humanos" que implica este tipo de evaluación[1] y concluyen que, a pesar de existir grandes diferencias entre distintos evaluadores y entre diferentes modelos, la clasificación de los sistemas participantes apenas cambia cuando se promedian los resultados obtenidos al trabajar con varios conjuntos de documentos, modelos y jueces. Señalan que, naturalmente, habrá diferencias en los resúmenes de documentos individuales pero que la única forma de mejorar la tecnología de resumen automático es mejorando los resultados promedio en este tipo de evaluaciones.

Lin y Hovy (2002b) también analizaron la influencia del factor humano en *DUC 2001* y estudiaron la posibilidad de sustituir este tipo de evaluaciones por métodos automáticos. Llegaron a las siguientes conclusiones[2]: (1) las evaluaciones humanas son "inestables", es decir, dos revisores pueden asignar puntuaciones distintas al comparar la misma sentencia con un modelo; (2) los distintos sistemas evaluados quedan separados en distintos "grupos de rendimiento" por lo que, a pesar de todo, los revisores humanos, tomados en conjunto, demuestran un criterio que permite establecer comparaciones entre sistemas; (3) es posible desarrollar un método automático que otorgue puntuaciones basándose en la coincidencia de *n*-gramas de palabras entre resumen y modelo; (4) la clasificación de los participantes obtenida mediante dicho sistema automático muestra una elevada correlación con la clasificación producida por los revisores humanos y (5) los autores de los modelos suelen construir sentencias nuevas por lo que la única forma en que un sistema de evaluación automática puede enfrentarse a estos problemas es empleando varios (probablemente muchos) modelos.

Posteriormente (Lin y Hovy 2003) estudiaron la posibilidad de utilizar *BLEU* (Papineni *et al.* 2002), una herramienta para la evaluación de sistemas de traducción automática, para la evaluación de resúmenes automáticos. *BLEU* emplea una media ponderada del número de *n*-gramas de palabras de longitud variable que coinciden entre una traducción automática y un modelo de traducción. Comprobaron que esta medida no siempre exhibía una correlación con las clasificaciones producidas por evaluadores humanos mientras que una medida basada en la coincidencia de unigramas (esto es, palabras aisladas) mostraba un mejor comportamiento y, en consecuencia, abría las puertas al desarrollo de métricas que pudiesen ser obtenidas de modo automático y, al mismo tiempo, garantizar una evaluación similar a la que podría realizar un revisor humano.

A raíz de estos trabajos se implementó un sistema de evaluación automático denominado *ROUGE* (Lin 2004a) que comenzó a utilizarse en *DUC 2004*. Esta herramienta permite calcular diversas medidas, principalmente *ROUGE-N*, *ROUGE-L* y *ROUGE-W*. La primera se basa en el número de *n*-gramas de palabras que coinciden entre un resumen candidato y uno o más modelos por lo que existen las medidas *ROUGE-1*, *-2*, *-3*, etc. no siendo habitual emplear más alla de los 4-gramas. *ROUGE-L* emplea la longitud de las subcadenas más largas que son comunes en el candidato y en el modelo mientras que *ROUGE-W* es una versión ponderada de *ROUGE-L* que además de la longitud de la subcadena valora la ausencia de "huecos" en la misma (véase Fig. 118).

---

[1] La variabilidad tanto entre revisores como entre los modelos construidos para realizar la evaluación.

[2] Estas conclusiones encuentran apoyo en otros autores, así van Halteren (2002) en relación a *DUC 2002* afirma *"no está claro si uno o dos extractos creados manualmente constituyen una referencia suficiente"*. Por su parte, Santos, Mohamed y Zhao (2004) también propusieron un sistema de evaluación automática aunque en su caso los resúmenes no eran comparados con ningún modelo sino con los propios documentos de partida.

1. Opposition leaders Prince Norodom **Ranariddh and Sam Rainsy**
2. **Ranariddh and Sam Rainsy** have charged that Hun Sen's victory in the elections was…
3. **Ranariddh and** his opposition ally, **Sam Rainsy**, refused to accept the election results
4. **Ranariddh and** former finance minister **Sam** Ram **Rainsy** have refused to enter into a coalition

Fig. 118 Diferencias entre las puntuaciones *ROUGE-L* y *ROUGE-W*.

Tomando la primera sentencia como referencia las tres siguientes tienen la misma puntuación *ROUGE-L* puesto que en los tres casos la longitud de la subcadena común más larga (en negrita) es la misma. No obstante, la "similitud" con la sentencia modelo es mayor en el segundo caso que en el tercero y en éste que en el cuarto debido a que las correspondientes subcadenas no son contiguas. La medida *ROUGE-W* permite capturar estas particularidades.

Lin (2004b) volvió a analizar el método de evaluación empleado en *DUC* a raíz de las críticas al mismo hechas por Nenkova y Passonneau (2004): *"las puntuaciones de DUC no pueden usarse para distinguir un buen resumen humano de uno malo; además, el método de DUC no es suficiente para diferenciar entre sistemas automáticos"*. Lin demuestra que dichas conclusiones no son correctas y que la metodología empleada en *DUC* es válida, en particular en lo que se refiere al número de modelos y muestras enviadas por cada participante. Además de esto, utilizando los datos de las ediciones de 2001, 2002 y 2003 concluye que las medidas obtenidas empleando *ROUGE* muestran una elevada correlación con las evaluaciones humanas hechas en *DUC* y puesto que éstas ofrecen resultados significativos concluye que no sólo la metodología de evaluación es válida sino que es posible llevarla a cabo de modo completamente automático.

En resumen, en la actualidad es posible evaluar de manera sistemática métodos de obtención de resúmenes (fundamentalmente extracción) a partir de textos de estilo periodístico. Sin embargo, aún quedan muchos aspectos a resolver en el campo del resumen automático como, por ejemplo, la necesidad de afrontar otros estilos más allá del periodístico o de adaptar los resúmenes a los propósitos específicos de los usuarios. Sin abandonar la evaluación intrínseca automática todo esto requeriría, sin duda, metodologías de evaluación extrínsecas basadas en tareas realistas (Spärck-Jones *et al.* 2004).

4 Resultados obtenidos por *blindLight*

A fin de analizar la viabilidad de la nueva técnica para la extracción automática de resúmenes se decidió utilizar los productos correspondientes a *DUC 2004*. Las razones fueron dos: por un lado se trataba de la primera campaña en que se había llevado a cabo una evaluación automática facilitando su reutilización y por otro, al tratarse de la edición más reciente, permitiría comparar la propuesta del autor con las técnicas más avanzadas disponibles.

En dicha edición se propusieron cinco tareas:

1. Resúmenes muy cortos, máximo 75 caracteres, a partir de un único documento.
2. Resúmenes cortos, máximo 665 caracteres, a partir de un conjunto de documentos.
3. Resúmenes muy cortos a partir de traducciones automáticas y manuales de árabe a inglés.
4. Resúmenes cortos a partir de un conjunto de traducciones automáticas y manuales de árabe a inglés.
5. Resúmens cortos creados a partir de un conjunto de documentos y "guiados" por consultas del tipo `"Who is X?"` (`"¿Quién es X?"`).

Las tareas 1 a 4 fueron evaluadas empleando únicamente *ROUGE* y la última mediante *SEE* (véase nota al pie en pág. 170), es decir, las cuatro primeras tareas se

evaluaron automáticamente y la quinta de manera manual. Para las tareas 1 y 2 se emplearon 50 conjuntos, alrededor de 500 documentos, pertenecientes a la colección *TDT*[1] en inglés. Para las tareas 3 y 4 se usaron 25 conjuntos, del orden de 250 documentos, pertenecientes a la misma colección en árabe y para la última tarea se utilizaron 50 conjuntos, aproximadamente 500 documentos, de la colección *TREC*[2] en inglés. En todos los casos los documentos eran noticias procedentes de agencias de prensa (p.ej. *Associated Press* y *New York Times* en el caso de la *TDT*). Para cada tarea se definieron unos métodos *baseline* que también participaron en la evaluación. En el caso de la primera tarea el resumen *baseline* consistía en los 75 primeros caracteres del documento y en la segunda se construía con los primeros 665 caracteres del documento más reciente.

Para la evaluación de *blindLight* como técnica de extracción de resúmenes se optó por realizar únicamente los experimentos relativos a las dos primeras tareas: resúmenes muy cortos o cortos de textos escritos (no traducidos) en inglés. No obstante, la resolución de ambas tareas requirió algunas modificaciones en la aplicación de la técnica. Por un lado, los resúmenes muy cortos podían ser simples listas de palabras clave (*NIST* 2004) por lo que parecía razonable añadir, además de la extracción de una única sentencia, la selección de palabras clave a los métodos de producción de resúmenes. Por otro, *blindLight* no contempla, a día de hoy, la extracción de resúmenes multidocumento por lo que para afrontar la segunda tarea fue necesario fusionar[3] las noticias de cada conjunto en un único documento para su resumen; naturalmente, ese no es el enfoque más adecuado (véase Fig. 119).

> **A fire turned a dance hall jammed with teen-age Halloween revelers into a deathtrap, killing at least 60 people and injuring about 180 in Sweden's second-largest city.** The building had just two exits, one of which was blocked by fire, city police technician Stephen Holmberg was quoted as saying by the Swedish news agency TT. Jamal Fawz, graf 19 pvs. **A fire turned a dance hall jammed with teen-age Halloween revelers into a deathtrap, killing 65 people and injuring 157 others in Sweden's second-largest city.** The fast-spreading fire that broke out just a few minutes before midnight Thursday gutted the building and left rescuers facing a hideous scene that local rescue service leader Lennart Olin likened to a "gas chamber".

Fig. 119 Ejemplo de resumen multidocumento obtenido con *blindLight*.

En este ejemplo puede apreciarse uno de los inconvenientes del sistema de resúmenes multi-documento descrito. Como se puede ver hay dos sentencias (la primera y la cuarta) prácticamente idénticas. Esto debería solucionarse en el futuro, sin embargo, no debería plantear excesivos inconvenientes puesto que podrían utilizarse las consabidas medidas Π y P para determinar qué sentencias se "solapan" y no deben, por tanto, formar parte del resumen final.

Así, para la construcción de resúmenes muy cortos con *blindLight* se plantearon los siguientes métodos:

- Extraer la sentencia con mayor significatividad media por caracter.
- Extraer la sentencia con mayor puntuación por fragmentos.
- Extraer la sentencia más relevante mediante la combinación de las puntuaciones por fragmentos y por significatividad media por carácter.

---

[1] *TDT – Topic Detection and Tracking* (Detección y Seguimiento de Temas) hace referencia a una iniciativa *DARPA* para el desarrollo de métodos para la detección y agrupamiento de material relativo a un tema específico extraído a partir de artículos o locuciones periodísticas tanto en inglés como en mandarín <http://www.nist.gov/speech/tests/tdt>.

[2] La colección de documentos desarrolladas para las *Text REtrieval Conferences* <http://trec.nist.gov>.

[3] El orden en que se unieron los documentos es aquel en el que aparecen dentro de su directorio; puesto que la fecha forma parte del nombre del archivo, en la mayor parte de los casos el orden fue cronológico a excepción de aquellos conjuntos que combinan noticias de distintas agencias (p.ej. APW19981028.0444 y NTY19981026.0292).

- Extraer las frases[1] (no sentencias) con mayor puntuación por palabras clave ordenándolas según su orden de aparición en el texto.
- Extraer una lista de palabras clave ordenadas por primera aparición en el texto. Dado que las "palabras clave" obtenidas por *blindLight* pueden ser frases el resumen podría contener palabras repetidas.
- Extraer una lista de palabras clave no repetidas.

Además, se experimentó con distintos tamaños de *n*-grama, de ventana (en el caso de los métodos que implican la extracción de palabras clave) y de estadísticos para el cálculo del peso de los *n*-gramas (información mutua, *SCP*, ganancia de información, etc.). Por otro lado, se probó un sistema de **compresión de sentencias**[2] trivial consistente en eliminar los *n*-gramas menos significativos hasta reducir la longitud de la sentencia extraída a 75 caracteres; en aquellos casos en que no se aplicó esta "compresión" simplemente se truncaba el resumen.

Los resultados detallados de la evaluación se presentan en un anexo mientras que las conclusiones a las que se ha llegado son las siguientes:

1. Los estadísticos más adecuados para la ponderación de los *n*-gramas de cara a la extracción de resúmenes son información mutua, *SCP* y Dice.
2. La utilización de *blindLight* para la construcción de resúmenes muy cortos (máximo 75 caracteres) no parece adecuada puesto que en todas las medidas a excepción de *ROUGE-1* el rendimiento es sustancialmente inferior a la media de participantes. En cambio, las palabras clave extraídas sí parecen ser relativamente útiles puesto que empleando *ROUGE-1* (unigramas) su rendimiento es ligeramente superior a la media.
3. El método de compresión de sentencias propuesto resulta contraproducente para los resúmenes muy cortos al ofrecer un rendimiento mucho peor que el simple truncamiento. Estos resultados coinciden con los de Lin (2003), por lo que antes de plantearse la eliminación del método de compresión habría que estudiar su aplicabilidad al texto de manera global y no local.
4. Debido a estos resultados es difícil afirmar qué método de los anteriormente mencionados resulta más útil para la construcción de resúmenes extractivos tan cortos. No obstante, a la luz de los resultados generales de *DUC 2004* en que el *baseline* (primeros 75 caracteres del documento) ofreció sistemáticamente resultados inapreciablemente peores que el mejor participante, inapreciablemente superiores a los 5 mejores participantes, sustancialmente mejores que la media de participantes y sólo fue superado de manera rotunda por los seres humanos parece necesario

[1] Recuérdese que por sentencia entendemos una oración gramatical mientras que una frase tan sólo es una secuencia de palabras dotadas de sentido pero que no forman oración. Por ejemplo, la sentencia "`Welcome to Wikipedia, the free-content encyclopedia that anyone can edit`" constaría de dos frases separadas por una coma.

[2] Mani, Gates y Bloerdom (1999) describen un sistema para la eliminación de frases innecesarias. Jing (2000) describe un sistema similar aunque más flexible (no se aplicaría a todas las sentencias extraídas) y general (se basa en distintas fuentes de conocimiento y no únicamente en heurísticos). Knight y Marcu (2000) desarrollaron un sistema para la compresión de sentencias empleando pares `<abstract, texto>` como datos de entrenamiento. Resulta interesante el estudio de Lin (2003) según el cual un resumen extractivo construido a partir de sentencias comprimidas no resulta necesariamente mejor, aún así *"existe potencial en la compresión de sentencias pero es necesario encontrar un mejor sistema de compresión que tenga en cuenta para la optimización aspectos globales entre distintas sentencias"*.

plantearse la utilidad de métodos tan complejos para extraer resúmenes de semejante longitud a partir de textos de estilo periodístico.

5. En en el caso de los resúmenes cortos (máximo 665 caracteres) sí hay un método preferible sobre los demás: la combinación de las puntuaciones obtenidas a partir de los fragmentos y de la significatividad media por carácter. Empleando dicho método, *blindLight* ha ofrecido rendimientos entre apreciable y sustancialmente superiores a la media de participantes en *DUC 2004* y, dependiendo de la medida empleada, entre sustancial y apreciablemente inferiores a los alcanzados por los 5 mejores participantes.
6. Por lo que respecta al tamaño de *n*-grama empleado parece ser preferible la utilización de 4-gramas sobre 3-gramas. No obstante, las diferencias no son apreciables para ninguna medida *ROUGE* a excepción de *ROUGE-3* y *ROUGE-4* en que resultan sustanciales.

En resumen, la utilización de *blindLight* como método de extracción de palabras clave sin tratar de construir un "titular" legible ofrece resultados muy similares a los de la mayor parte de tecnologías disponibles y, al igual que éstas, se encuentra muy lejos de alcanzar los resultados de un sistema tan sencillo como extraer los primeros caracteres de un artículo. Estaría por determinar si este fenómeno se limita a textos periodísticos o es generalizable a otros estilos.

En cuanto a resúmenes de mayor longitud pero igualmente cortos (máximo 665 caracteres) se puede afirmar que *blindLight* es, cuando menos, apreciablemente mejor que muchas de las tecnologías disponibles. De las diversas configuraciones admitidas por la técnica propuesta por el autor la que ofrece de manera sistemática los mejores resultados es aquella que emplea 4-gramas, información mutua como método de ponderación de los mismos y determina la relevancia de las sentencias a extraer combinando la puntuación obtenida a partir de los fragmentos y de la significatividad media por carácter.

Ejemplo de resumen humano

Lebanon's Parliament voted the country's top military man, Gen. Emile Lahoud, president. Lahoud, who promises to clean up a graft-riddled government, is popular and is backed by powerful Syria. It is unclear, though, whether Prime Minister Hariri, in office since 1992 and credited with the country's economic recovery, will continue to head the cabinet. 31 of 128 legislators chose not to support him, leaving it to the president to name the next prime minister. Consequently, Hariri withdrew his candidacy, claiming the president acted unconstitutionally when he accepted the mandate to name a prime minister. Hariri's administration was plagued by nepotism.

Resultado del mejor participante en *DUC 2004* *(ROUGE-L)*

The commander of Lebanon's army will become the country's next president after winning the crucial backing of Syria, the powerbroker in Lebanon. According to the constitution, a president must begin his term by appointing the prime minister and a Cabinet through a decree issued after consulting with Parliament members. Sources close to the prime minister, whose Cabinet has been in care-taker capacity since last Tuesday's swearing in of Emile Lahoud as president, said Hariri turned down the invitation from Lahoud to select a new Cabinet. Hariri is credited with restoring economic confidence and stabilizing the national currency.

Resumen *blindLight*

The 128-member legislature is expected to meet Thursday to elect Lahoud after outgoing President Elias Hrawi signs the constitutional amendment. Under a formula aimed at preventing the recurrence of the 1975-90 civil war, power in Lebanon is shared equally by a Maronite Christian president, a Sunni Muslim prime minister and a Shiite Parliament speaker. Prime Minister Rafik Hariri has declined an informal invitation from Lebanon's new president to form the next government, sparking a political crisis in this country as it rebuilds from its devastating civil war. Lahoud had been expected to issue a presidential decree last week asking Hariri to form the next government after the president polled members of the 128-seat Parliament on their choice for prime minister.

Fig. 120 Comparación de uno de los peores (*ROUGE-L*) resúmenes *blindLight* con los correspondientes a un ser humano y al mejor participante en *DUC 2004*.

Ejemplo de resumen humano

In a move widely viewed as an effort to placate the far right as he moves to withdraw from more West Bank land, Israeli Prime Minister Netanyahu named hardliner Ariel Sharon foreign minister and chief peace negotiator. Sharon, a military leader with legendary victories in the 1967 and 1973 Mideast wars, is infamous in the Arab world as the defense minister in the 1982 invasion of Lebanon during which Lebanese Christian militiamen, Israeli allies, slaughtered hundreds of unarmed Palestinians. His appointment as lead negotiator was denounced as a "disaster" in the Lebanese press and "a bullet of mercy to the peace process" in a Syrian paper.

Resultado del mejor participante en *DUC 2004* *(ROUGE-L)*

Sharon is the general who led Israel's 1982 invasion of Lebanon, and a former housing minister who strengthened Jewish settlement in territories Israel captured from Syria, Jordan and Egypt in the 1967 Mideast war. Ariel Sharon has a law degree, and fancies himself a farmer. 1996: Named infrastructure minister in Netanyahu government. Prime Minister Benjamin Netanyahu named Sharon foreign minister on Friday, effectively putting the hard-liner in charge of negotiating Israel's final borders with the Palestinians. Palestinians also were expressing growing unease over the naming of hawkish former Israeli general Ariel Sharon as Netanyahu's foreign minister.

Resumen *blindLight*

Just days before heading to the United States for critical negotiations with Palestinian leaders, Prime Minister Benjamin Netanyahu jolted the Middle East peace effort with the appointment of Ariel Sharon as Israeli foreign minister. Ariel Sharon's appointment as the Israeli foreign minister serves as "the bullet of mercy" for the Middle East peace process, an official Syrian newspaper said Saturday. Prime Minister Benjamin Netanyahu named Sharon foreign minister on Friday, effectively putting the hard-liner in charge of negotiating Israel's final borders with the Palestinians. An Israeli tribunal looking into the invasion found him indirectly responsible for the massacre of hundreds of Palestinian refugees by Christian Lebanese militiamen at two Beirut camps.

Fig. 121 Comparación de uno de los mejores (*ROUGE-L*) resúmenes *blindLight* con los correspondientes a un ser humano y al mejor participante en *DUC 2004*.

4.1 Variabilidad de los resultados entre distintos idiomas

Parece claro que el resumen automático no debe limitarse a unos pocos idiomas:

> *Si un sistema de resumen automático emplea métodos tomados de la recuperación de información y, por tanto, independientes del idioma o si los métodos para un idioma específico pueden simplificarse y adaptarse a otros se podrá adaptar rápidamente un sistema de extracción de resúmenes en un idioma para que funcione con otros.* (Hovy 1999)

Dentro de este entorno multilingüe hay dos escenarios que reciben mucha atención[1]. Por un lado, estarían aquellos sistemas que deben resumir en un único idioma "objetivo" documentos escritos en una variedad de idiomas "fuente". Por ejemplo, Evans y Klavans (2003) o Evans, Klavans y McKeown (2004) generan resúmenes en inglés a partir de noticias escritas en inglés o traducidas automáticamente[2] al inglés desde otros lenguajes. En segundo lugar se encontrarían aquellos en los que deben producirse resúmenes en un idioma arbitrario a partir de documentos escritos en dicho idioma u otro diferente, por ejemplo, el proyecto *MUSI* (Busemann 2001) o (Lenci *et al.* 2002) que a partir de artículos de medicina escritos en inglés o italiano obtiene resúmenes en alemán y francés, o el trabajo desarrollado por Gawronska (2002) que permite el resumen de noticias escritas en inglés al danés, sueco y polaco. En este segundo escenario el enfoque extractivo no parece ser el preferido por los investigadores y se tiende a la utilización de representaciones abstractas de los documentos que permiten la generación del resumen final en el idioma seleccionado (Busemann 2001), (Gawronska 2002).

No obstante, y como paso previo al desarrollo de tales sistemas, resulta interesante el estudio de técnicas estadísticas de resumen automático que sean fácilmente adaptables a distintos idiomas y ofrezcan resultados consistentes con independencia del lenguaje a

[1] En particular en la Unión Europea.

[2] Las tareas tercera y cuarta de *DUC 2004* se correspondían con este escenario, los sistemas participantes debían resumir tanto documentos escritos en inglés como traducidos al inglés desde el árabe (*NIST* 2004).

resumir. En este sentido cabría señalar el trabajo realizado con dos sistemas ya mencionados, ambos extractivos y basados en técnicas independientes del idioma: *SUMMARIST* (Hovy y Lin 1997) y *MEAD* (Radev, Blair-Gondensohn y Zhang 2001). El primero de ellos fue adaptado con facilidad para su aplicación al bahasa Indonesia[1] (Lin 1999) mientras que el segundo se ha extendido para su funcionamiento en chino e inglés y en teoría con cualquier otro lenguaje natural (Radev *et al.* 2004). Otro sistema de resumen multilingüe, según esta interpretación, fue *MINDS* (Cowie *et al.* 1998) que permitía la obtención de resúmenes a partir de documentos escritos en coreano, español, japonés o ruso.

Por último, hay que destacar el trabajo llevado a cabo por Radev *et al.* (2002) para la evaluación de sistemas de resumen automático en entornos multilingües. Estos investigadores han publicado[2] un *corpus* paralelo de textos en chino e inglés junto con resúmenes creados a partir de cada documento individual y de distintos subconjuntos del *corpus*.

Por su propia naturaleza *blindLight* debería ser aplicable a distintos idiomas sin mayores problemas. No obstante, la cuestión no es su aplicabilidad sino su variabilidad, esto es, la influencia que tiene el idioma en que está escrito un documento en el momento de extraer su resumen. Dicho de otro modo, dadas dos traducciones literales de un texto ¿son también traducciones literales los resúmenes extraídos?

Sin lugar a dudas, hubiera resultado muy interesante la utilización del *corpus* paralelo antes mencionado para la evaluación de la nueva técnica y en el futuro se llevará a cabo tal evaluación. No obstante, el autor ha realizado por el momento una pequeñísima "prueba de concepto" para verificar hasta qué punto los resultados de la técnica que propone son invariables respecto a los distintos idiomas.

Para ello se seleccionaron cinco[3] notas de prensa de la Comisión Europea[4] que disponían de versiones en alemán, danés, francés, húngaro e inglés. Todos estos idiomas, a excepción del húngaro que es urálico, son indoeuropeos de los cuales todos, a excepción del francés que es romance, son germánicos siendo, a su vez, el danés un idioma germánico nórdico y alemán e inglés germánicos occidentales. La longitud media de los documentos empleados es de 2800 caracteres pero hay diferencias sustanciales entre los distintos idiomas. Por ejemplo, el francés es el más verboso con un promedio de 3.073 caracteres por documento frente al inglés con 2.506 caracteres. Por último, aun cuando las traducciones son literales hasta donde ha podido comprobar el autor, debido al uso de la puntuación el número de sentencias puede variar ligeramente de un idioma a otro. Por ese motivo los documentos originales fueron "corregidos" para garantizar que el número de sentencias era idéntico en todas las traducciones y permitir así la utilización del coeficiente de Spearman[5] para comprobar la correlación entre los distintos resúmenes.

---

[1] La variedad de malayo hablada en ese país.

[2] `http://www.clsp.jhu.edu/ws2001/groups/asmd`

[3] "*French journalists win first EU 'For Diversity. Against Discrimination' Award*", *"The European Union on your doorstep: new generation of information relays launched"*, *"Europeans want policy makers to consider the environment as important as economic and social policies"*, *"European Commission launches investigations into sharp surge in Chinese textiles imports"* y *"Post tsunami: the Commission reinforces its disaster response capacity"*.

[4] `http://europa.eu.int/comm/press_room/index_en.htm`

[5] Véase página 75 y posteriores.

El método empleado en este pequeño experimento es similar al empleado por Radev *et al.* (2002) para obtener resúmenes extractivos "manuales" para cada documento del *corpus* paralelo chino-inglés que, recordemos, estaba alineado a nivel de sentencia. En su caso recurrieron a revisores que para cada sentencia asignaron una puntuación de 0 a 10 en función de su utilidad para un resumen extractivo. Una vez hecho esto, para obtener un resumen extractivo "manual" de un documento tan sólo era necesario indicar el porcentaje de compresión y seleccionar aquellas sentencias más útiles hasta completar la longitud deseada; el resumen equivalente en el otro idioma se construía mediante las sentencias homólogas. De este modo, fue posible disponer de cientos de resúmenes elaborados según un criterio humano y garantizando que los resúmenes en ambos idiomas eran traducciones literales.

En este caso también se dispone de un mínimo conjunto de documentos alineado a nivel de sentencia (después de corregir la puntuación en algún caso concreto) y de un evaluador automático: *blindLight*. Lo que se desea saber es hasta qué punto la técnica es invariable frente al idioma, es decir, en qué medida se acerca al ideal según el cual siempre determinaría la sentencia más relevante con independencia del idioma. Para ello, se procedió a procesar cada documento mediante la nueva técnica obteniendo para cada una de sus sentencias la significatividad media por carácter.

A partir de aquí resulta inmediato elaborar para cada documento una lista de sentencias ordenadas por relevancia decreciente, listas que pueden ser comparadas mediante el coeficiente de Spearman para comprobar su correlación. Recordemos que dicho coeficiente varía en el intervalo [-1, 1] donde -1 significa que hay una correlación negativa perfecta, 0 la ausencia de correlación y 1 la existencia de una correlación positiva perfecta.

Por ejemplo, si al comparar la lista de sentencias ordenadas para un documento d_i en los lenguajes L_1 y L_2 se obtuviese -1 significaría que las sentencias más importantes en un idioma serían sistemáticamente elegidas como las menos relevantes en el otro y viceversa; si el coeficiente fuese 0 no habría ninguna correlación y en caso de obtener la unidad la técnica funcionaría de un modo "ideal" puesto que habría otorgado a cada sentencia la misma relevancia con independencia del idioma.

En la Tabla 24 se muestran los resultados obtenidos al analizar la correlación existente entre las ordenaciones producidas por *blindLight* para cada idioma. Como se puede ver, en todos los casos la correlación es superior a 0,5 y ciertamente es elevada aunque lejos de ser ideal. Por otro lado, el grado de correlación parece venir marcado por el "parentesco" de los distintos idiomas[1], lo cual era de esperar, y en cierta medida por la longitud del documento pues en los más largos la correlación es menor[2]. La misma prueba fue llevada a cabo para el sistema *MEAD* (Radev, Blair-Goldensohn y Zhang 2001) y como se puede comprobar (véase Tabla 25) el comportamiento de ambos sistemas es muy similar.

En resumen, la técnica *blindLight* es una herramienta útil para la extracción de resúmenes automáticos a partir de texto libre escrito en cualquier lenguaje natural. Su aplicación para la obtención de resúmenes muy cortos (máximo 75 caracteres) no parece adecuada pero, a la luz de las evaluaciones realizadas hasta la fecha, muy pocas técnicas consiguen superar la eficacia de un método tan sencillo como extraer los primeros

[1] Por ejemplo, alemán, danés e inglés presentan los valores más elevados mientras que el húngaro es el que obtiene los menores valores en todos los casos.

[2] Aunque el autor no ha llevado a cabo ningún experimento en relación con la influencia del tamaño de los documentos es muy probable que la previa segmentación del documento en pasajes, mediante técnicas análogas a *TextTiling* (Hearst 1994), influya positivamente en la calidad del resumen final.

caracteres del documento. Por lo que respecta a la extracción de resúmenes cortos (máximo 665 caracteres) su rendimiento es superior a muchas de las técnicas más avanzadas disponibles y aunque aún no se ha implementado un verdadero sistema de resumen multidocumento su desarrollo resulta natural debido a las características de la técnica para la detección de similitudes "semánticas". Por lo que respecta a la invariabilidad de los resultados respecto al idioma a resumir las pruebas preliminares indican que es bastante elevada aunque la naturaleza del idioma influye de manera importante. No obstante, con base en la experiencia previa, el autor tiene confianza en que empleando marcos de evaluación multilingüe o monolingüe en idiomas distintos al inglés será posible alcanzar buenos resultados en cualquier lenguaje natural.

EN/FR	EN/DE	EN/DA	EN/HU	FR/DE	FR/DA	FR/HU	DE/DA	DE/HU	DA/HU
0,50	0,68	0,77	0,81	0,63	0,71	0,35	0,90	0,58	0,55
0,72	0,74	0,73	0,61	0,67	0,78	0,67	0,81	0,55	0,52
0,36	0,43	0,52	0,15	0,42	0,59	0,63	0,71	0,12	0,20
0,60	0,83	0,80	0,72	0,64	0,68	0,61	0,82	0,67	0,77
0,43	0,89	0,93	0,93	0,71	0,46	0,54	0,82	0,93	0,79
0,52	**0,71**	**0,75**	**0,64**	**0,61**	**0,64**	**0,56**	**0,81**	**0,57**	**0,57**

Tabla 24. Coeficientes de correlación de Spearman entre los resúmenes obtenidos empleando información mutua y 3-gramas. La última fila es el valor medio.

EN/FR	EN/DE	EN/DA	EN/HU	FR/DE	FR/DA	FR/HU	DE/DA	DE/HU	DA/HU
0,49	0,71	0,86	0,54	0,30	0,55	0,55	0,53	0,20	0,56
0,48	0,37	0,46	0,84	0,60	0,81	0,46	0,48	0,44	0,47
0,53	0,67	0,49	0,82	0,41	0,41	0,55	0,24	0,75	0,48
0,56	0,68	0,81	0,45	0,63	0,63	0,41	0,79	0,53	0,41
0,89	0,94	-0,26	0,94	0,94	0,09	0,83	-0,20	0,89	-0,09
0,59	**0,67**	**0,47**	**0,72**	**0,58**	**0,50**	**0,56**	**0,37**	**0,56**	**0,37**

Tabla 25. Resultados de la misma prueba para el sistema *MEAD*.

La Comisión ha adoptado hoy propuestas relativas a un paquete de medidas destinadas a reforzar la capacidad de respuesta de la Unión Europea en caso de catástrofes. Estas medidas se destinan a financiar nuevos equipos especializados en materia de planificación para agilizar el suministro eficaz de ayuda a largo plazo; a reforzar la capacidad de la Unión de facilitar equipos de expertos civiles y de equipo y a suministrar ayuda humanitaria. La Comunicación adoptada hoy también presenta un informe detallado sobre la utilización de los 450 millones de euros anunciados por la UE tras la catástrofe del tsunami. Las propuestas adoptadas hoy constituyen la contribución de la Comisión al plan de acción tras el tsunami propuesto por la Presidencia luxemburguesa el 31 de enero.

«Vistas las situaciones anteriores y nuestra capacidad de responder inmediatamente ante la catástrofe del tsunami, la Comisión propone ahora medidas que nos ayudarán, en el futuro, a contribuir de forma rápida y eficaz a las tareas de reconstrucción tras una catástrofe» ha declarado la Comisaria de Relaciones Exteriores y Política de Vecindad, Benita Ferrero-Waldner, que propone dichas medidas conjuntamente con los Comisarios Michel y Dimas.

Stavros Dimas, Comisario Europeo responsable de Protección Civil ha dicho: «Nuestra reacción ante el Tsunami ha demostrado el claro valor añadido que la dimensión europea aporta a la asistencia en materia de protección civil. Las propuestas de hoy hacen avanzar un paso más al Mecanismo actual... Tomadas en su conjunto, permitirán disponer de un instrumento que garantiza una reacción europea eficaz ante futuras catástrofes».

The Commission has today adopted proposals for a package of measures to reinforce the European Union's disaster response capacity. The package will: fund new specialist planning teams to speed up the effective delivery of long term aid; reinforce the Union's capacity to provide specialised civil expertise units and equipment; and strengthen the Union's capacity to deliver humanitarian aid. The Communication adopted today also provides a detailed progress report of how the 450 million Euro pledged by the EU after the tsunami disaster is being spent. The proposals agreed today are the Commission's contribution to the post-tsunami Action Plan proposed by the Luxembourg Presidency on 31st January.

"Against our background and success to respond immediately to the Tsunami disaster the Commission proposes now measures that will help us to respond swiftly and effectively to post crisis reconstruction in the future" said Commissioner for External Assistance and European Neighbourhood Policy, Benita Ferrero-Waldner, who proposed the steps with Commissioners Michel, and Dimas.

Stavros Dimas, the European Commissioner responsible for Civil Protection, said: "The response to the Tsunami demonstrated the clear added value that the European dimension brings to civil protection assistance. The proposals made today take the existing Mechanism one step further. Taken together they will result in an instrument that guarantees an effective European reaction to future disasters."

La Commission a approuvé ce jour plusieurs propositions relatives à un train de mesures destinées à renforcer la capacité de réaction de l'Union européenne en cas de catastrophes. Il s'agit des mesures suivantes: financement accordé pour la mise en place d'équipes de spécialistes en matière de planification pour accélérer la fourniture efficace d'une aide à long terme et renforcement de la capacité de l'Union à mettre à disposition des équipes d'experts civils et du matériel et à effectuer des opérations d'aide humanitaire. La Communication adoptée ce jour fournit également un rapport détaillé sur l'utilisation de la contribution de 450 millions d'euros annoncée par l'Union européenne au lendemain du tsunami. Les propositions approuvées aujourd'hui constituent la contribution de la Commission au plan d'action après-tsunami proposé le 31 janvier dernier par la Présidence luxembourgeoise.

"Au regard de nos expériences antérieures et de notre capacité à réagir sans délai après la survenue du tsunami, la Commission propose maintenant des mesures qui nous permettront, dans l'avenir, de contribuer rapidement et en toute efficacité aux travaux de reconstruction faisant suite à des crises", a déclaré Benita Ferrero-Waldner, commissaire chargée des relations extérieures et de la politique européenne de voisinage, qui a présenté les mesures en question conjointement avec les commissaires Michel et Dimas.

"Notre réaction lors du tsunami témoigne clairement de la valeur ajoutée que la dimension européenne confère à l'assistance en matière de protection civile. Les mesures proposées aujourd'hui constituent une nouvelle avancée du mécanisme existant et elles permettront de disposer d'un instrument garantissant une réaction européenne efficace lors de prochaines catastrophes" a ajouté Stavros Dimas, commissaire européen chargé de la protection civile.

Die Kommission hat heute Vorschläge für ein Maßnahmenpaket angenommen, das die Katastrophenabwehrkapazitäten der Europäischen Union stärken soll, indem neue spezialisierte Planungsteams zur Beschleunigung der wirksamen Erbringung langfristiger Hilfe finanziert werden, die Kapazitäten der EU für die Bereitstellung spezialisierter zivilen Expertenteams und Ausrüstung verstärkt werden und die Kapazitäten der EU für die Erbringung humanitärer Hilfe ausgebaut werden. Die heute angenommene Mitteilung enthält außerdem einen ausführlichen Fortschrittsbericht darüber, wie die von der EU nach der Tsunami-Katastrophe bereitgestellten 450 Mio. EUR eingesetzt werden. Die heute gebilligten Vorschläge stellen den Beitrag der Kommission zum dem Aktionsplan dar, den der luxemburgische Vorsitz am 31. Januar infolge des Tsunami vorgelegt hatte.

"Vor dem Hintergrund des Erfolgs unserer unmittelbaren Reaktion auf die Tsunami-Katastrophe schlägt die Kommission nun Maßnahmen vor, die uns in die Lage versetzen werden, künftig rasch und wirksam zu reagieren, wenn es um Wiederaufbaumaßnahmen nach Krisen geht", kommentierte die für die Außenhilfe und die Europäische Nachbarschaftspolitik zuständige Kommissarin Benita Ferrero-Waldner, die die Maßnahmen gemeinsam mit den Kommissaren Michel und Dimas vorgestellt hat.

Der für den Katastrophenschutz zuständige Kommissar Stavros Dimas äußerte sich wie folgt: "Die Reaktion auf den Tsunami hat den deutlichen Mehrwert gezeigt, den die europäische Dimension für die Katastrophenhilfe erbringt. Die heute unterbreiteten Vorschläge gehen einen Schritt weiter als das bestehende Verfahren. Gemeinsam werden sie ein Instrument bilden, das eine effiziente europäische Reaktion auf künftige Katastrophen ermöglicht."

I dag vedtog Kommissionen en række forslag som led i en pakke af foranstaltninger for at styrke EU's katastrofeberedskab. Med pakken finansieres nye eksperthold, hvis planlægning skal sikre, at den langsigtede bistand bliver effektiv, EU's muligheder for at tilbyde specialiserede civile ekspertenheder og udstyr forbedres, og EU's muligheder for at yde humanitær bistand øges. Meddelelsen, som blev vedtaget i dag, rummer også en detaljeret statusrapport med oplysninger om, hvad de 450 mio. EUR, som EU har stillet til rådighed efter flodbølge-katastrofen, bruges til. Forslagene, der blev vedtaget i dag, er Kommissionens bidrag til den handlingsplan, det Luxembourgske formandskab foreslog den 31. januar 2005 som opfølgning på flodbølgekatastrofen.

"I lyset af den hurtige og vellykkede indsats i forbindelse med flodbølgekatastrofen foreslår Kommissionen nu foranstaltninger, som sætter os i stand til fremover hurtigt og effektivt at yde en genopbygningsindsats efter en nødsituation", sagde kommissæren for eksterne forbindelser og den europæiske naboskabspolitik, Benita Ferrero-Waldner, der sammen med kommissær Louis Michel og Stavros Dimas foreslog dette tiltag.

Stavros Dimas, der er kommissær med ansvar for civilbeskyttelse, udtalte: "Indsatsen i forbindelse med flodbølgen viste, at den europæiske dimension skaber en klar merværdi i forbindelse med civilbeskyttelsesbistand. Den eksisterende ordning udvikles yderligere med de forslag, som er fremsat i dag. De vil samlet set munde ud i et beredskab, som garanterer en effektiv europæisk indsats i forbindelse med katastrofer i fremtiden."

A Bizottság ma javaslatokat fogadott el az Európai Unió katasztrófaelhárító képességének fokozására hivatott intézkedéscsomagot illetően. A csomag a hosszú távú segítségnyújtás hatékony felgyorsítása érdekében biztosítja az új szakértői tervezőcsoportok működéséhez szükséges anyagi fedezetet; fokozza az Unió azon képességét, hogy speciális civil szakértői csoportokat és felszerelést biztosítson; elősegíti, hogy az Unió nagyobb részt vállalhasson a humanitárius segítségnyújtásban. A ma elfogadott közlemény ezenkívül részletesen beszámol az EU által a cunami-katasztrófát követően felajánlott 450 millió euró felhasználásáról is. A javaslatok elfogadásával a Bizottság ahhoz a cselekvési tervhez kíván hozzájárulni, melyet a luxemburgi elnökség január 31-én, a cunamit követően terjesztett elő.

„Meglévő lehetőségeinkre és a cunami-katasztrófára adott gyors válaszintézkedéseinkre alapozva a Bizottság olyan intézkedéseket javasol, amelyeknek köszönhetően a jövőben gyorsan és hatékonyan reagálhatunk a válságokat követő újjáépítési szükségletekre" – nyilatkozta Benita Ferrero-Waldner, az EU külkapcsolatokért és európai szomszédsági politikáért felelős biztosa, aki Louis Michel és Stavros Dimas európai biztossal közösen javasolja ezeket az intézkedéseket.

Stavros Dimas, a polgári védelemért felelős biztos kijelentette: „A cunamira adott válaszlépéseink bebizonyították, mekkora többletet ad az európai dimenzió a polgári védelem által nyújtott segítséghez. A ma beterjesztett javaslatoknak köszönhetően újabb lépéssel viszik előre a meglévő mechanizmust és együttesen olyan eszköz létrejöttét eredményezik, amely garantálja, hogy Európa hatékonyan reagál a jövőbeni katasztrófákra."

Fig. 122 Resúmenes de aproximadamente el 25% de la última nota de prensa en español, inglés, francés, alemán, danés y húngaro (se ha conservado la puntuación original y el número de sentencias varía).

CAPÍTULO 8

CONCLUSIONES Y TRABAJO FUTURO

El tema central de este trabajo es la "sobrecarga de información" y la forma de afrontarla. El autor está interesado en una forma particular de información: texto libre escrito en cualquier lenguaje natural; y en un entorno específico en que se produce dicha sobrecarga: los medios *online*. Esta sobrecarga de información es un problema antiguo que se ha visto acrecentado por el progreso tecnológico: por un lado, el abaratamiento de los soportes físicos permite almacenar textos sin tener que plantearse la eliminación de ningún documento no importa lo obsoleto o inútil que sea (véase Fig. 123) y, por otro, la disponibilidad de estándares de comunicación, transporte, formateo de documentos, etc. que han permitido que prácticamente cualquier usuario pueda convertirse en fuente de información.

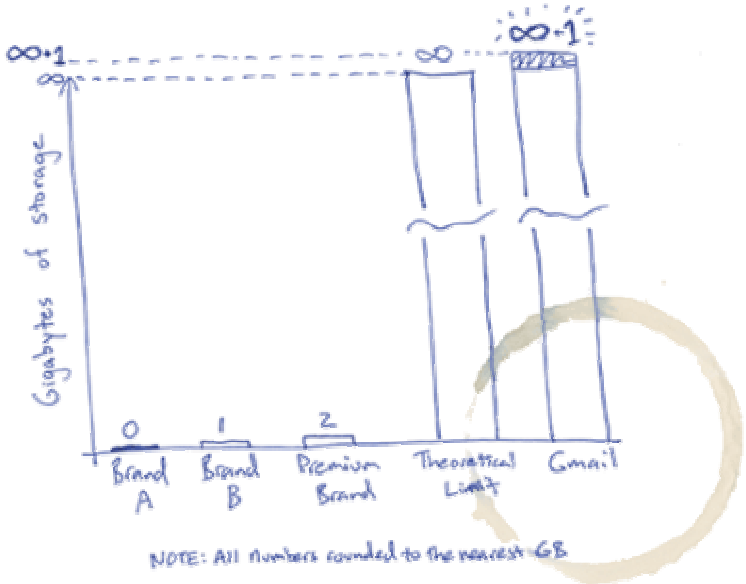


Fig. 123 Una "broma seria" de *Google* que ofrece a los usuarios de la solución de correo electrónico *GMail* una cuota creciente y presuntamente ilimitada a partir de 2GB.

El autor realizó una propuesta para afrontar dicho problema: la Web Cooperativa que se sustenta sobre tres puntos:

1. La utilización de conceptos, generados automáticamente, como alternativa intermedia entre las ontologías y las palabras clave.
2. La clasificación de documentos en una taxonomía a partir de tales conceptos.
3. La cooperación entre usuarios, en realidad, entre agentes que actúan en representación de los usuarios y que no requieren su participación explícita.

El primer punto venía motivado por los inconvenientes de la utilización de palabras clave para la formulación de las "necesidades de información" por parte de los usuarios y la dificultad para desarrollar ontologías que diesen soporte a cualquier consulta concebible. El autor planteó los conceptos como entidades más abstractas y, por tanto, con mayor carga semántica que las palabras clave pero que, al mismo tiempo, pudiesen ser obtenidos de manera automática. Una posible forma de construir tales conceptos sería mediante "agrupaciones débiles" de términos relacionados y una tecnología que podría llevar a cabo esta tarea es la semántica latente aunque también podría utilizarse la técnica presentada como parte central de este trabajo.

Frente a la Web Semántica que requiere la utilización de un "marcado ontológico" la Web Cooperativa propone la utilización de una semántica más sencilla[1]: simples categorías obtenidas de manera automática a las que los distintos documentos podrían asociarse sin necesidad de emplear ningún tipo de etiquetas. Para ello el autor sugiere utilizar tan sólo el texto plano de los documentos, argumentando que éstos pueden considerarse individuos de una población mayor y que, del mismo modo en que es posible clasificar y categorizar de manera automática a los seres vivos mediante su código genético, es posible adaptar algoritmos empleados en biología computacional al campo de la clasificación de documentos.

Por último, con la Web Cooperativa se pretende aprovechar el conocimiento experimental que obtienen los usuarios al explorar la Web actual y que se desaprovecha en gran medida. Para ello se pretende utilizar técnicas que permitan obtener de los usuarios, de forma no intrusiva y transparente, información sobre la relevancia de los distintos documentos. Así, cada usuario de la Web Cooperativa dispondría de un agente con dos objetivos: aprender de su "maestros" y recuperar información para él.

Por tanto, el objetivo último de la Web Cooperativa sería, por un lado, dotar de una cierta semántica a la Web mediante técnicas automáticas y explotar la experiencia diaria de los usuarios en provecho de los mismos: facilitándoles la ejecución de consultas adaptadas a sus necesidades y ofreciéndoles, sin precisar de consulta alguna, información relevante.

Así pues, la Web Cooperativa involucra muy diversos aspectos: tratamiento de lenguaje natural, evaluación implícita de documentos, agentes *software*, interacción persona-ordenador, usabilidad o privacidad. Por otro lado, sus objetivos son muy ambiciosos y exceden con creces no sólo los fines de un trabajo como esta disertación sino, sobre todo, las capacidades de un único investigador. Por esa razón se plantearon una serie de cuestiones previas para determinar que subconjunto de las mismas constituían, a un

[1] Desde hace algún tiempo existe una *meme* que también defiende el uso de etiquetas más sencillas pero aún dotadas de semántica: "folksonomía". Este término, del inglés *folksonomy* = *folk* + *taxonomy*, hace referencia a la categorización colaborativa de documentos mediante etiquetas (palabras clave) escogidas libremente por los usuarios. Aún nos encontramos en una fase muy incipiente como para predecir el desarrollo de estos métodos pero será necesario prestarles atención.

tiempo, un problema interesante y practicable. Las preguntas finalmente escogidas fueron las siguientes:

- ¿Es posible clasificar textos libres empleando métodos tomados de la biología computacional?
- ¿Es posible obtener un "pseudo-ADN" a partir de texto escrito en un lenguaje natural?
- Si existiera ese pseudo-ADN, ¿sería posible combinarlo, mutarlo o construir "algo" a partir del mismo?
- ¿Se debe suponer que idiomas distintos constituyen "bioquímicas" diferentes?
- ¿Cómo se daría el salto desde ese pseudo-ADN a los conceptos?
- ¿En qué forma podría un agente realizar búsquedas eficientes sobre una taxonomía de documentos construida a partir de ese pseudo-ADN?

Así, para la realización de este trabajo se prescindió de lo que serían las "capas superiores" de la Web Cooperativa y se delimitó mejor el problema a resolver:

> *La cantidad de texto no estructurado disponible en la Web seguirá aumentando y, a pesar de sus inconvenientes, el método preferido por la mayor parte de usuarios para recuperar información continuarán siendo las consultas formuladas en lenguajes naturales. En ambos casos (publicación y consulta) será inevitable un uso generalmente ambiguo de los distintos idiomas y la presencia de errores tipográficos, ortográficos o gramaticales.*

De este modo se encuadró el problema dentro del campo de procesamiento de lenguaje natural por medios estadísticos y se formuló la siguiente tesis:

> *Se puede obtener para los distintos n-gramas, g_i, de un texto escrito en cualquier idioma una medida de su significatividad, s_i, distinta de la frecuencia relativa de aparición de los mismos en el texto, f_i, pero calculable a partir de la misma. Esta métrica de la significatividad intradocumental de los n-gramas permite asociar a cada documento, d_i, un único vector, v_i, susceptible de comparación con cualquier otro vector obtenido del mismo modo aun cuando sus respectivas longitudes puedan diferir. Puesto que tales vectores almacenan ciertos aspectos de la semántica subyacente a los textos originales, el mayor o menor grado de similitud entre los mismos constituye un indicador de su nivel de relación conceptual, facilitando la clasificación y categorización de documentos, así como la recuperación de información. Asimismo, cada vector individual es capaz de transformar el texto original a partir del cual fue obtenido dando lugar a secuencias de palabras clave y resúmenes automáticos.*
>
> Que se resumía de este modo:
>
> *Una única técnica sencilla, basada en el uso de vectores de n-gramas de longitud variable, independiente del idioma y aplicable a diversas tareas de tratamiento de lenguaje natural con resultados similares a los de otros métodos 'ad hoc' es viable.*

En resumen, aunque el trasfondo de este trabajo es la sobrecarga de información el problema de que realmente se ocupa es el del **procesamiento de texto natural por medios estadísticos y en condiciones "extremas"**: multilingüismo, gran número de documentos, textos ambiguos, no estructurados y con ruido (errores tipográficos, ortográficos o gramaticales).

Este problema, especialmente notable en los entornos *online*, puede desglosarse en una serie de tareas: asignación de documentos a categorías conocidas (**categorización**),

agrupación de documentos con características similares (**clasificación o *clustering***), **recuperación de información** y destilación de información (p.ej. **resumen automático**).

No es necesario decir que existen diversas técnicas capaces de afrontar una o más de las tareas anteriores. Sin embargo el autor planteó que una única técnica bastaba para resolver todas las tareas de manera adecuada verificando, además, las siguientes características: independencia del idioma, utilización de métodos puramente estadísticos y alta tolerancia al ruido.

Dicha técnica, propuesta por el autor y denominada *blindLight*, es una técnica parcialmente bioinspirada puesto que parte del concepto de "ADN documental" que estaría formado por una secuencia de genes, pares constituidos por un *n*-grama de caracteres y la significatividad de dicho *n*-grama dentro del texto del documento.

La principal diferencia entre esta propuesta y otras también basadas en la idea de un "genoma documental" radica en que no se pretende emplear éste únicamente para categorizarlo o clasificarlo sino que, al igual que el ADN de los seres vivos, este "ADN documental" puede combinarse entre sí además de "activarse" produciendo un resultado diferente del texto original.

Las razones para emplear *n*-gramas de caracteres que pueden "saltar" entre palabras son varias: (1) facilitan un trato "igualitario" a todos los lenguajes aun cuando no utilicen separadores de palabras (p.ej. el chino o el japonés), (2) garantizan una tolerancia elevada al ruido, (3) permiten obviar el uso de ciertos algoritmos (p.ej. los de *stemming*) y ofrecen un rendimiento adecuado incluso en idiomas muy complejos (p.ej. el finés).

Por lo que respecta al cálculo de la significatividad de cada *n*-grama en el documento se pueden emplear toda una serie de estadísticos (p.ej. información mutua, Dice, χ^2, probabilidad condicional simétrica, etc.) que no requieren la utilización de un contexto en que situar al documento (algo necesario si se utilizase, por ejemplo, *tf*idf*).

Por otro lado, en tanto que cadenas, en particular de inspiración biológica, parecía claro que sería posible aplicar algoritmos tomados de la biología computacional y llevar a cabo la clasificación automática de documentos representados de este modo. Sin embargo, era precisa una solución más general que permitiese determinar la similitud entre dos cadenas de este "ADN documental".

Antes de definir esa medida de asociación se describió un proceso de "hibridación", o mejor, intersección entre cadenas de pseudo-ADN de tal modo que a partir de dos de tales cadenas se obtuviese una tercera. En términos biológicos se puede afirmar que cuanto mayor sea la longitud de la cadena híbrida más elevado resulta el grado de parentesco entre las dos cadenas originales. De modo análogo, *blindLight* define una operación de intersección de cadenas de "ADN documental" y, en consecuencia, establece dos medidas asimétricas denominadas Π (Pi) y P (Ro) que vinculan la significatividad total de la cadena intersección o híbrida con la de cada uno de los dos progenitores.

Dependiendo de la longitud de los documentos a comparar las significatividades totales de ambos y de la cadena intersección pueden ser muy distintas y, en consecuencia, los valores de Π y P muy diferentes. No obstante, el hecho de que sean independientes permite su combinación lineal de diversas formas y su adaptación a las diversas necesidades de cada tarea.

Uno de los aspectos que pueden resultar más controvertidos de la tesis del autor es la afirmación de que los vectores de *n*-gramas empleados para representar este pseudo-ADN

son capaces de almacenar aspectos semánticos subyacentes al texto original, en particular, si se tiene en cuenta la pretensión de que la técnica es válida para cualquier tipo de lenguaje natural. No obstante, una serie de experimentos relativos a la clasificación de traducciones literales de un conjunto de documentos en español, inglés, francés, finés, holandés, hebreo y japonés así como el resumen de textos en inglés, alemán, francés, danés y húngaro concluyeron que ***blindLight***, aun cuando no sea total y absolutamente independiente respecto al idioma, **muestra un comportamiento extremadamente consistente entre lenguajes muy diferentes** por lo que se puede afirmar que, efectivamente, este "ADN documental" sí almacena ciertos aspectos semánticos del texto. Una vez garantizado este aspecto de la técnica se describió su aplicación a cada una de las tareas anteriormente descritas: clasificación, categorización, recuperación de información y destilación de información (en particular extracción de resúmenes y palabras clave).

Para la primera tarea, la **clasificación automática de documentos**, se presentaron dos algoritmos basados en *blindLight* (uno incremental y otro no incremental) y se comparó la nueva técnica propuesta con otros métodos, resultando que ***blindLight* ofrece un rendimiento similar al de ciertas técnicas** (p.ej. mapas auto-organizativos) **y mejor que el de otras como los métodos particionales y jerárquicos**. También se llevó a cabo un experimento relativo a la clasificación genética de lenguajes naturales empleando textos de 14 idiomas europeos y transcripciones fonéticas de 9 idiomas. Los resultados de ambas clasificaciones no sólo fueron coherentes entre sí sino también con las teorías lingüísticas vigentes. A raíz de estas experiencias se concluyó que, en efecto, era posible emplear *blindLight* como técnica de clasificación automática con resultados análogos a los de métodos específicos.

Posteriormente se describió la aplicación de la técnica propuesta por el autor a la **categorización de documentos** y se llevaron a cabo una serie de experimentos relativos a identificación de idiomas, autoría de documentos, filtrado de correo no deseado así como una prueba con las colecciones *Reuters 21578* y *OHSUMED*. A la luz de tales experimentos se puede concluir:

1. La utilización de *blindLight* como sistema para la identificación de idiomas proporciona unos resultados muy similares (y bajo ciertas condiciones de longitud del texto y ruido superiores) a técnicas reconocidas aun empleando apenas 10KB de información para cada idioma.
2. La aplicación de *blindLight* como filtro de *spam* requiere mejoras pero como simple experimento de categorización parece sugerir un rendimiento similar al de los clasificadores Bayesianos y *MBL* (*Memory Based Learning*).
3. Las pruebas estandarizadas indican que *blindLight* ofrece resultados análogos a clasificadores Bayesianos, Rocchio y árboles de decisión, apreciablemente inferiores a *k*-vecinos y sustancialmente inferiores a las *SVM* (*Support Vector Machines*).

En resumen, ***blindLight* no alcanza los resultados de las *SVM* pero es similar a otras técnicas empleadas con frecuencia y aceptadas como adecuadas** (p.ej. clasificadores Bayesianos).

Para la evaluación de *blindLight* como técnica de **recuperación de información** se experimentó con las colecciones *CACM* y *CISI* y se tomó parte en la edición de 2004 del *CLEF*. Es necesario decir que los resultados obtenidos en ambos casos fueron claramente **inferiores a los de las técnicas tradicionales** por lo que la afirmación respecto a la viabilidad de esta técnica para su utilización en *IR* queda, por el momento, en suspenso. No obstante, es necesario señalar una serie de aspectos alentadores y que sugieren que, en el

futuro, tal vez sea posible situar a *blindLight* al mismo nivel que otras técnicas consolidadas. En primer lugar, aunque las técnicas tradicionales de recuperación de información superan a la del autor, otras nuevas y consideradas "prometedoras" (como el indexado por semántica latente) ofrecen resultados muy similares; por otro lado, las medidas de similitud entre consultas y documentos pueden mejorarse (tal vez empleando programación genética) y uno de los elementos empleados en *CLEF'04* (el sistema de pseudo-traducción) aún estaba en una fase preliminar.

Por lo que respecta a la **extracción de resúmenes** y palabras clave se describió el modo en que se puede utilizar el "ADN documental" para segmentar el texto plano del documento en fragmentos de máxima significatividad empleando un procedimiento inspirado en la síntesis de las proteínas. Este método permite obtener una información muy valiosa para determinar las sentencias más relevantes del texto y construir así un resumen extractivo. Para evaluar este enfoque se emplearon los datos de la edición 2004 de *DUC* (*Document Understanding Conferences*) obteniendo unos resultados muy alentadores: al extraer resúmenes cortos (máximo 665 caracteres) a partir de un conjunto de documentos, ***blindLight* resultó ser superior a muchas de las técnicas más avanzadas disponibles**, aunque aún está lejos de alcanzar a los mejores sistemas existentes en la actualidad.

Por lo que respecta al desarrollo futuro de este trabajo existen varias líneas interesantes:

1. Adaptar el sistema de extracción de resúmenes a entornos multidocumento.
2. Continuar el desarrollo del sistema de pseudo-traducción.
3. Analizar la posible integración de los dos sistemas anteriores.
4. Emplear programación genética para la obtención de nuevas medidas de similitud entre documentos y consultas en el sistema *IR*.
5. Estudiar la posible integración de medidas basadas en la complejidad de Kolmogorov.
6. Estudiar la utilización de fragmentos de significatividad máxima como términos de indexado en el sistemas *IR*.

En conclusión, el autor ha presentado una **técnica novedosa** para el **procesamiento de lenguaje natural** por medios puramente estadísticos. Dicha técnica es aplicable a **múltiples idiomas** ofreciendo resultados consistentes en todos ellos, muestra una adecuada **tolerancia al ruido** y resulta **apta para tareas de clasificación, categorización y extracción de resúmenes**. Además, parece **potencialmente útil** para la **recuperación de información** en entornos **multilingües** aunque en este campo aún no se ha progresado lo suficiente.

GLOSARIO

Agente *software*

Un agente *software* es un programa autocontenido que opera como parte de un entorno, es capaz de tomar decisiones sobre la base de su percepción de dicho entorno y puede ejecutar acciones para alcanzar una serie de objetivos. Los agentes *software* actúan en representación de otros actores (agentes o, más comunmente, usuarios) y requieren nula o muy poca participación de los mismos para su funcionamiento.

Agrupamiento (*clustering*)

Véase **clasificación automática**.

Agrupamiento incremental/no incremental

Dos tipos de **clasificación automática**. El agrupamiento incremental trabaja sobre elementos aislados y parte del supuesto de que es posible considerarlos de uno en uno asignándolos a algún grupo ya disponible; por esta razón está especialmente indicada para conjuntos muy grandes. Por su parte, el agrupamiento no incremental opera sobre todo el conjunto de elementos, generalmente requiere comparar todos los elementos entre sí y, en consecuencia, es apta sólo para conjuntos relativamente pequeños.

Árbol de decisión

Una técnica de **categorización automática** en la que los nodos del árbol se corresponden con variables, los arcos con valores para las mismas y las hojas con las categorías predichas basándose en los valores de las variables que se encuentran en el recorrido entre la raíz del árbol y la hoja.

Automatic summarization

Véase **resumen automático**.

Autoridad

Según Kleinberg (1998) una autoridad es una página web fuertemente enlazada, o lo que es lo mismo, referenciada.

blindLight

Técnica estadística de **procesamiento de lenguaje natural** que establece métodos para la representación vectorial de textos escritos en cualquier idioma así como para la comparación de dichos vectores permitiendo el desarrollo de algoritmos de **clasificación**, **categorización**, **recuperación de información** y **resumen automático**. Estos vectores emplean *n*-gramas de caracteres a los que se asocia un valor de **significatividad** a partir tan sólo de los contenidos del documento original.

Boosting hypothesis

Hipótesis planteada por Kearns (1988) acerca de la posibilidad de construir un categorizador eficiente con una cadena de "categorizadores débiles" (aquellos cuya regla de decisión es sólo ligeramente mejor que una decisión tomada al azar). Se basa en el modelo de aprendizaje automático ***PAC***, según el cual el aprendizaje exitoso equivale a la minimización del error de la hipótesis (o regla de decisión).

Cadena léxica

Una cadena léxica es una secuencia de palabras semánticamente relacionadas que aparecen en un texto y que pueden ser adyacentes o encontrarse dispersas a lo largo del documento. Para encontrar dichas cadenas léxicas en un texto genérico es necesario utilizar recursos como *WordNet* que proporcionan la información necesaria sobre las posibles relaciones entre distintas palabras.

Categorización automática

Término que hace referencia a una amplia variedad de técnicas que tienen como objetivo asignar a un objeto dado una o más categorías (o etiquetas) de un conjunto predefinido. La categorización automática de documentos se realiza a partir del texto de los mismos y requiere una fase previa de entrenamiento durante la cual el categorizador es enfrentado con unos pocos ejemplos de las distintas categorías que debe reconocer.

Categorización mediante *boosting*

Una técnica de **categorización automática** basada en la denominada ***boosting hypothesis***. Se trata de un meta-algoritmo puesto que parte de la utilización de distintas técnicas de categorización. A diferencia de la **categorización mediante comités**, en el *boosting* los distintos categorizadores trabajan por etapas: (1) un categorizador se entrena sobre una parte del conjunto de entrenamiento y se prueba sobre el resto del conjunto; (2) aquellos documentos del conjunto de entrenamiento que clasifique mal, junto con algunos otros de su subconjunto de entrenamiento original, se utilizan para entrenar otro categorizador que, de este modo, "aprende" casos más "difíciles"; (3) este esquema se repite *n* veces.

Categorización mediante comités

Técnica de **categorización automática** basada en el uso simultáneo de varios categorizadores (un comité) que emiten un voto para cada elemento a categorizar. El resultado de dicha votación establece la categoría o categorías finalmente asignadas.

Categorizador bayesiano (*naïve Bayes*)

Técnica de **categorización automática** basada en el teorema de Bayes de probabilidad condicionada y que supone una total independencia entre las características que definen cada objeto. El apelativo de *naïve* se debe a lo irreal de esta suposición. La base de estos categorizadores es la siguiente: (1) a partir del conjunto de entrenamiento se puede establecer una probabilidad *a priori* para cada categoría y la probabilidad de cada característica condicionada para cada categoría; (2) para categorizar un objeto (definido por los valores de las características) basta con usar los datos anteriores para calcular la probabilidad de cada categoría condicionada a las características observadas en el objeto.

Centroide

Dado un conjunto de puntos multidimensionales (véase **modelo vectorial**) el centroide es aquel punto que tiene como coordenadas la media de los valores en cada dimensión.

Clasificación automática (agrupamiento o *clustering*)

Término que hace referencia a una amplia variedad de técnicas que, dentro de un conjunto de elementos, permiten identificar grupos que exhiben características similares (véase **similitud**). Para ello pueden dividir de manera iterativa el conjunto original en subconjuntos o comenzar por los elementos aislados e ir agrupando los más próximos. Así mismo, existe la posibilidad de operar sobre todo el conjunto simultaneamente o de manera paulativa (véase **Agrupamiento incremental/no incremental**).

Clustering (agrupamiento)

Véase **clasificación automática**.

Compresión de sentencias (*sentence compression*)

Técnica relacionada con el **resumen automático**. Según Knight y Marcu (2000) la compresión de sentencias tiene como finalidad conservar la información más relevante de una sentencia reescribiéndola en una forma más corta. El grado de sofisticación de estas técnicas es muy variable. Según Lin (2003) un **resumen extractivo** construido a partir de sentencias comprimidas no resulta necesariamente mejor que un resumen extractivo de la misma longitud sin compresión; aún así afirma que *"existe potencial en la compresión de sentencias pero es necesario encontrar un mejor sistema de compresión que tenga en cuenta para la optimización aspectos globales entre distintas sentencias"*.

Concentrador (*hub*)

Según Kleinberg (1998) un *hub* (o concentrador) es una página web que contiene enlaces a varias **autoridades**.

Consulta

Una consulta es la manifestación escrita de una necesidad de información formulada por un usuario a un sistema de **recuperación de información**.

Consulta informativa

Según Broder (2002) aquella **consulta** que un usuario formula en un buscador web para obtener algún tipo de información que supone está disponible en una o más páginas web (p.ej. `degenerative disc disease` o `muscle aches during pregnancy`).

Consulta navegacional

Según Broder (2002) aquella **consulta** que un usuario formula en un buscador web para llegar a un sitio web en particular (p.ej. `hotmail`, `renfe` o `universidad de oviedo`)

Consulta transaccional

Según Broder (2002) aquella consulta que un usuario formula en un buscador web para alcanzar un sitio web en el que llevar a cabo algún tipo de actividad (p.ej. `hotmail`, `weather` o `maps`)

Corpus

En lingüística un *corpus* es una colección, generalmente muy grande, de documentos (típicamente textos o habla transcrita) que muestran el uso real de una lengua natural. Un *corpus* puede contener muestras de un único lenguaje (*corpus* monolingüe) o de varios lenguajes (*corpus* multilingüe). En el caso de los *corpora* multilingües pueden ser comparables (el número de documentos y términos son similares para todos los idiomas y la temática es homogénea) o paralelos (los documentos son traducciones

literales y se dispone de información sobre su "alineación" a nivel de documento, párrafo o sentencia).

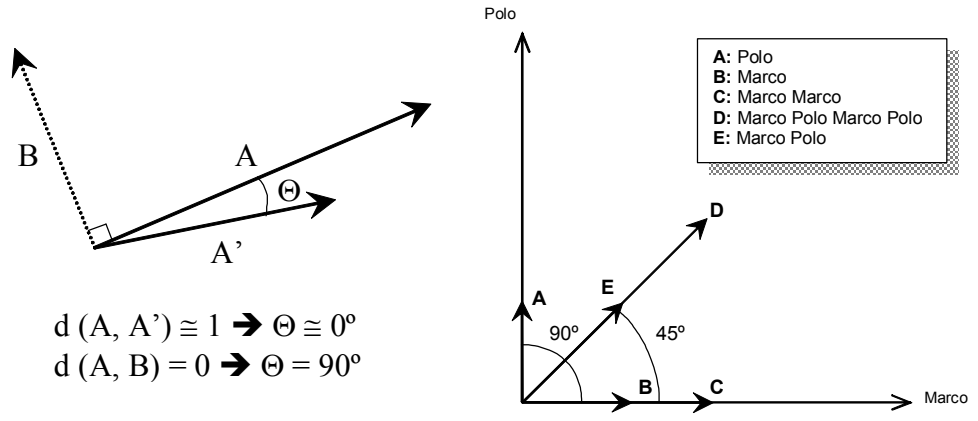
Coseno, función del (*cosine similarity*)

La función del coseno es una medida de **similitud** comunmente empleada en el **modelo vectorial**. Se calcula mediante la siguiente ecuación en la que n es el número de términos (dimensiones del espacio vectorial) y q_i y d_i son, respectivamente, el i-ésimo término de los documentos q y d.

$$\frac{\sum_{i=1}^{n} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}}$$

Puesto que en el modelo vectorial no se usan generalmente pesos negativos esta función ya es una similitud normalizada que, además, admite una interpretación geométrica sencilla: cuanto más próximo a 1 esté el valor obtenido más cercano a 0º será el ángulo formado por los vectores y, en consecuencia, más similares serán éstos; por el contrario, valores próximos a 0 implicará que los vectores son ortogonales (la máxima separación posible en un espacio vectorial en el que todos los términos toman valores positivos).

En la siguiente figura se muestran cinco documentos representados en un espacio vectorial de dos dimensiones así como los ángulos entre los vectores de algunas parejas ilustrativas. Obsérvese que para la función del coseno (y en general para cualquier medida de similitud) no se puede afirmar que una similitud total (en este caso un ángulo de 0º como el que forman *D* y E) implique la identidad entre ambos documentos. Compárense con los obtenidos con una medida de **disimilitud**.



Destilación de información

La destilación de información está relacionada en cierta medida con la **recuperación de información**; en el contexto de este trabajo hace referencia a técnicas como la respuesta de preguntas (*question answering*) o el **resumen automático** (*automatic summarization*).

Disimilitud

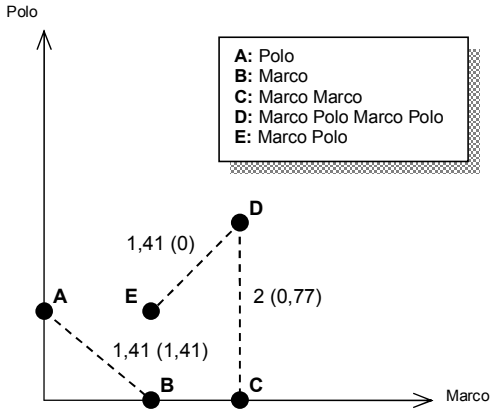
La disimilitud mide la discrepancia entre dos objetos a partir de sus características. Puesto que el **modelo vectorial** define un espacio multidimensional es posible determinar la disimilitud entre dos documentos i y j simplemente calculando la distancia entre ambos:

$$d_{ij} = \sqrt[p]{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^p}$$

En la ecuación anterior, n es el número de términos o dimensiones del espacio, k es el término k-ésimo de cada documento y p es el orden de la distancia. Esta distancia se denominada de Minkowski y para p=1 equivale a la distancia Manhattan, para ese mismo valor pero datos binarios a la distancia de Hamming y para p=2 a la distancia euclídea.

No resulta demasiado adecuado emplear directamente la distancia puesto que se ve muy influida por el tamaño de los documentos. Esto puede paliarse en parte normalizando los vectores pero aun así no se dispondrá de una disimilitud normalizada (que varía entre 0 y 1). No obstante, a partir de las mismas características se puede calcular de manera directa una medida de **similitud** que, por otra parte, puede convertirse de manera trivial en una disimilitud normalizada si fuese necesario.

En la siguiente figura se muestran cinco documentos representados en un espacio vectorial de dos dimensiones así como las distancias euclídeas entre algunas parejas ilustrativas (entre paréntesis la distancia euclídea para los vectores normalizados). Obsérvese las diferencias entre los resultados obtenidos antes y después de la normalización de los vectores; compárense con los obtenidos con una medida de similitud como la **función del coseno**.



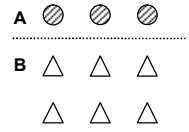
Entropía

La entropía es una medida para la evaluación de soluciones de **agrupamiento** (que producen grupos de elementos) mediante la comparación con una clasificación previa (que proporciona una serie de clases). Dado un grupo S_r de tamaño n_r su entropía se define como:

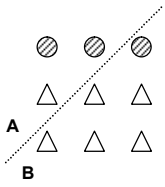
$$E(S_r) = -\sum_{i=1}^{q} p_{ir} \log p_{ir} = -\sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r}$$

donde q es el número de clases y n^i_r es el número de documentos de la clase i-ésima que fueron asignados al grupo r-ésimo. La entropía de la clasificación final se define como la suma de las entropías de todos los grupos ponderadas de acuerdo a su tamaño, es decir:

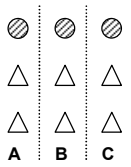
$$Entropia = \sum_{r=1}^{k} \frac{n_r}{n} E(S_r)$$



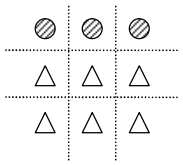
$$E(A) = -\left(\tfrac{3}{3} \cdot \log \tfrac{3}{3} + \tfrac{0}{3} \cdot \log \tfrac{0}{3}\right) = -(1 \cdot \log 1 + 0 \cdot \log 0) = 0$$
$$E(B) = -\left(\tfrac{0}{6} \cdot \log \tfrac{0}{6} + \tfrac{6}{6} \cdot \log \tfrac{6}{6}\right) = -(0 \cdot \log 0 + 1 \cdot \log 1) = 0$$
$$E = 0$$



$$E(A) = -\left(\tfrac{2}{3} \cdot \log \tfrac{2}{3} + \tfrac{1}{3} \cdot \log \tfrac{1}{3}\right) = -(-0{,}117 - 0{,}159) = 0{,}276$$
$$E(B) = -\left(\tfrac{1}{6} \cdot \log \tfrac{1}{6} + \tfrac{5}{6} \cdot \log \tfrac{5}{6}\right) = -(-0{,}130 - 0{,}066) = 0{,}196$$
$$E = \tfrac{3}{9} \cdot 0{,}276 + \tfrac{6}{9} \cdot 0{,}196 = 0{,}223$$



$$E(A) = E(B) = E(C) =$$
$$= -\left(\tfrac{1}{3} \cdot \log \tfrac{1}{3} + \tfrac{2}{3} \cdot \log \tfrac{2}{3}\right) = -(-0{,}159 - 0{,}117) = 0{,}276$$
$$E = 0{,}276$$



$$E(A) = E(B) = \ldots = E(H) = E(I) =$$
$$= -\left(\tfrac{1}{1} \cdot \log \tfrac{1}{1} + \tfrac{0}{1} \cdot \log \tfrac{0}{1}\right) = -(1 \cdot \log 0 + 0 \cdot \log 0) = 0$$
$$E = 0$$

Cuanto mayor es la semejanza de la solución obtenida y la clasificación externa menor es la entropía de dicha solución. El valor mínimo posible es 0 que supondría una clasificación idéntica a la externa o bien una solución trivial consistente en la división del conjunto de elementos en grupos formados por un único *ítem* (véase último caso en la figura).

Exhaustividad (*recall*)

La exhaustividad es una medida de la "calidad" de un sistema de **recuperación de información**. Se trata de la fracción del total de documentos relevantes que son obtenidos por un sistema *IR*. La exhaustividad máxima es, por definición, 1 y conlleva la recuperación de todos los documentos relevantes existentes. Esta medida por sí sola no es suficiente para caracterizar a un sistema *IR* puesto que se puede garantizar trivialmente una exhaustividad total recuperando todos los documentos de la colección para cualquier **consulta**. Por esa razón al evaluar un sistema siempre se calcula, además de la exhaustividad, su **precisión**. Ambas pueden combinarse en un solo valor: la **medida *F***. Véase además **curva de precisión-exhaustividad**.

Fallout (tasa o índice de irrelevancia, índice de fallos)

Una medida del rendimiento de un sistema de **recuperación de información** relacionada con la **precisión** y la **exhaustividad** aunque no tan utilizada como éstas. Se trata de la proporción de documentos no relevantes en la colección que se ofrecen como resultados de una consulta. Así pues, informa sobre la rapidez con que la precisión disminuye al aumentar la exhaustividad (o lo que es lo mismo, el número de documentos retornados por consulta).

Filtrado colaborativo

El filtrado colaborativo es una técnica que permite a un sistema sugerir a cada usuario en particular una selección de nuevos elementos sobre la base de sus preferencias en el pasado y de las valoraciones que, de dichos elementos, han hecho otros usuarios del sistema que coinciden, en mayor o menor medida, con las preferencias del usuario

original. Un ejemplo típico es el servicio de *Amazón* (`http://www.amazon.com`) *"Customers who bought this book also bought…"* ("Los clientes que compraron este libro también compraron…")

Folksonomía (*folksonomy*)

Término procedente del inglés *folksonomy* = *folk* + *taxonomy* que hace referencia a la categorización colaborativa de documentos (en general páginas web) mediante etiquetas (palabras clave) escogidas libremente por los usuarios. Un ejemplo del incipiente uso de folksonomías puede encontrarse en `http://del.icio.us`, un sitio web donde los usuarios almacenan sus enlaces favoritos etiquetándolos y pudiendo descubrir, a través de la exploración de dichas etiquetas, nuevos sitios web potencialmente relevantes para sus intereses.

Hub

Véase **concentrador**.

idf (*inverse document frequency*)

Método para ponderar los términos de un documento en un sistema de **recuperación de información**. Fue propuesto por Karen Spärck-Jones (1972) y se basa en la idea de que un término es tanto más informativo y, en consecuencia, importante cuanto menor es el número de documentos que lo contienen. Es decir, el peso de un término es inversamente proporcional al número de documentos que lo emplean. La expresión habitualmente empleada para el cálculo del valor *idf* es la siguiente:

$$w = -\log \frac{n}{N}$$

Donde w es el peso del término, n es el número de documentos que contienen dicho término y N es el total de documentos de la colección. Con frecuencia este método de ponderación se combina para formar el denominado ***tf*idf***.

Intersección Ω

En el contexto de la técnica ***blindLight*** se trata de un operador que permite la combinación de dos vectores documentales en un nuevo vector intersección. Dicho vector contiene los *n*-gramas que aparecen en los dos vectores originales y para cada uno de dichos *n*-gramas se asocia como significatividad el valor mínimo del par.

IR (*information retrieval*)

Véase **recuperación de información**.

Macropromediar y micropromediar (*macroaverage* vs. *microaverage*)

Dos formas de obtener resultados promedio al evaluar sistemas de **categorización automática** según Lewis (1991). Dado un conjunto de D documentos y una serie de K categorías un categorizador toma $D \cdot K$ decisiones que pueden ser evaluadas individualmente. A fin de ofrecer un único valor promedio puede obtenerse la precisión para cada categoría y posteriormente calcular su media o bien tomar todas las decisiones como un único conjunto. En el primer caso se habla de *macroaveraging* y en el segundo de *microaveraging*.

La diferencia entre una y otra medida es simple: en el caso de *microaveraging* tiene más influencia el resultado global (esto es, el número total de categorizaciones correctas) frente a las diferencias entre categorías (puede haber diferencias notables entre los resultados obtenidos para cada categoría) mientras que en el caso de *macroaveraging* influyen más las diferencias entre categorías que los resultados tomados en su

conjunto, es decir, se "premiaría" al categorizador que obtiene resultados similares en todas las categorías. En función del tipo de aplicación debe decidirse qué tipo de comportamiento es preferible y emplear un tipo u otro de promedio para la evaluación de los resultados.

A continuación se muestra un pequeño ejemplo. Supongamos una colección de documentos en distintos idiomas: 1000 en inglés (EN), 600 en español (ES), 300 en portugués (PT), 100 en alemán (DE) y 20 en francés (FR). Las categorías serían naturalmente los nombres de los idiomas y la lengua en que está escrito cada documento no se conoce *a priori*, es decir, deben categorizarse los documentos de acuerdo a su idioma. Supongamos que los resultados obtenidos al categorizar automáticamente dicha colección son los siguientes:

Categoría	Resultados		Precisión
EN	950 (EN)	50 (DE)	0,95
ES	600 (ES)	100 (PT)	0,86
PT	150 (PT)	10 (FR)	0,94
DE	50 (DE)	50 (EN)	0,50
FR	10 (FR)	50 (PT)	0,17
	1760	260	**Macro: 0,68**
	Micro: 0,87		

En este caso el valor macropromedio es de 0,68 y el micropromedio de 0,87. Es decir, un 87% del total de documentos fueron clasificados correctamente; sin embargo, puesto que el valor macropromediado es mucho menor puede afirmarse que este sistema se comporta de manera sustancialmente diferente frente a las distintas categorías.

Mapas Auto-organizativos (*Self-Organizing Maps* o *SOM*)

Véase ***Self-Organizing Maps***.

Máquinas de Vectores Soporte

Véase ***Support Vector Machines***.

Medida *F*

La medida *F* fue propuesta por van Rijsbergen (1979) como una medida única para evaluar la calidad de un sistema de **recuperación de información**. Esta medida combina en un solo valor la **precisión** y **exhaustividad** de un sistema:

$$F = \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)}$$

En la ecuación anterior la precisión es *P* y la exhaustividad *R* (*recall*). Se suele utilizar un valor de α=½ por lo que la ecuación generalmente empleada para calcular la medida *F* es la siguiente:

$$F = 2\frac{P \cdot R}{P + R}$$

Medoide

Aquel elemento de un conjunto de puntos (véase **modelo vectorial**) que está más próximo a su **centroide**.

Modelo vectorial (*vector space model*)

Modelo propuesto por Salton y Lesk (1965) consistente en la representación de un conjunto de documentos como puntos en un entorno *T*-dimensional siendo *T* el número de términos distintos en el conjunto. Los términos son generalmente palabras o raíces o lemas de palabras. Cada documento será pues un vector de pesos, siendo estos nulos si el término no aparece en el documento y no nulos si el documento lo contiene; en este caso pueden usarse distintos métodos de ponderación, típicamente ***tf*idf***. Dada la naturaleza algebraica del modelo es posible definir **distancias** (y **similitudes**) entre los documentos siendo habitual el uso de la **función del coseno**.

Naïve Bayes

Véase **categorizador bayesiano**.

N-grama

Un *n*-grama es una secuencia de *n* elementos, palabras o caracteres, extraídos de un texto de forma no necesariamente correlativa. En el contexto de este trabajo se entiende por *n*-grama una secuencia de *n* caracteres contiguos que puede contener blancos y, por tanto, estar formado por segmentos de varias palabras consecutivas. Por ejemplo, los cinco primeros 3-gramas de esta definición serían `Un_`, `n_n`, `_n-`, `n-g` y `-gr` (se han reemplazado los blancos por guiones bajos).

NLP (*Natural Language Processing*)

Véase **procesamiento de lenguaje natural**.

Ontología

Una ontología, en un contexto informático, es según Gruber (1993) *"la especificación de una conceptualización. Esto es, una descripción de los conceptos y relaciones que pueden existir para un agente o una comunidad de agentes"* . La **Web Semántica** se basa en el uso intensivo de ontologías definidas como *"un documento o fichero que define formalmente las relaciones entre términos; una ontología típica para la Web consta de una taxonomía y de un conjunto de reglas de inferencia"* (Berners-Lee, Hendler y Lassila 2001, p.4).

PAC (*Probably Approximately Correct*)

Modelo matemático de aprendizaje automático (aplicable, por tanto, a la **categorización automática**) que establece la equivalencia entre aprendizaje exitoso y la minimización del error de una hipótesis (o regla de decisión) obtenida a partir de ejemplos de entrenamiento tomados al azar. Las hipótesis (o reglas) aprendidas son aproximadas pues fallan para una fracción de ejemplares establecida de manera arbitraria.

PageRank

PageRank (Page *et al.* 1998) hace referencia a un algoritmo para el cálculo del "prestigio" de una página web así como al valor calculado por dicho algoritmo. La técnica se basa en la idea de citación o referencia según la cual un documento muy citado (o lo que es lo mismo, enlazado) será más prestigioso que otro menos citado o no citado en absoluto (véase **autoridad**). Sin embargo, a diferencia de otros métodos, *PageRank* no considera por igual todos los enlaces recibidos por un documento sino en función del valor numérico (también *PageRank*) del documento del que parte el enlace. De este modo el "prestigio" o autoridad se propaga mediante los enlaces de unos documentos a otros: el *PageRank* de una página se divide por el número de enlaces de salida y se "transfiere" a los documentos enlazados. Así, documentos que reciben muchos enlaces aunque de poco valor serán muy relevantes y documentos que reciben pocos enlaces pero desde páginas con *PageRank* elevado serán igualmente

importantes. *PageRank* es uno de los métodos que emplea el buscador web *Google* (`http://www.google.com/`) para determinar la relevancia de los documentos que satisfacen una **consulta**.

Palabras vacías

Véase ***stop words***.

Pasaje

Según Salton *et al.* (1996) cada uno de los *"fragmentos de texto [en que puede dividirse un documento] que exhiben consistencia interna y que pueden distinguirse del resto de texto circundante"*.

Pi (Π)

En el contexto de la técnica ***blindLight*** se define Π como el cociente de la **significatividad total** del vector **intersección** de una **consulta** y un documento entre la significatividad total del vector consulta. Π revela en qué medida la consulta queda "satisfecha" por la intersección entre ésta y un documento resultante. Esta medida puede combinarse con **Ro (P)** para construir distintas medidas de **similitud** adaptadas a las necesidades de las diferentes aplicaciones.

Precisión

La precisión es una medida de la "calidad" de un sistema de **recuperación de información**. Se trata de la fracción de documentos obtenidos por un sistema *IR* que son relevantes. La precisión máxima es, por definición, 1 y supone que todos los documentos recuperados son relevantes. Esta medida por sí sola no es suficiente para caracterizar a un sistema *IR* puesto que si no se retorna ningún documento la precisión sería 1 ya que no hay ningún resultado irrelevante. Por ello al evaluar un sistema siempre se calcula, además de la precisión, su **exhaustividad**. Ambas pueden combinarse en un solo valor: la **medida *F***. Véase además **precisión en *k***, **precisión media**, **precisión interpolada** y **curva de precisión-exhaustividad**.

Precisión en *k*

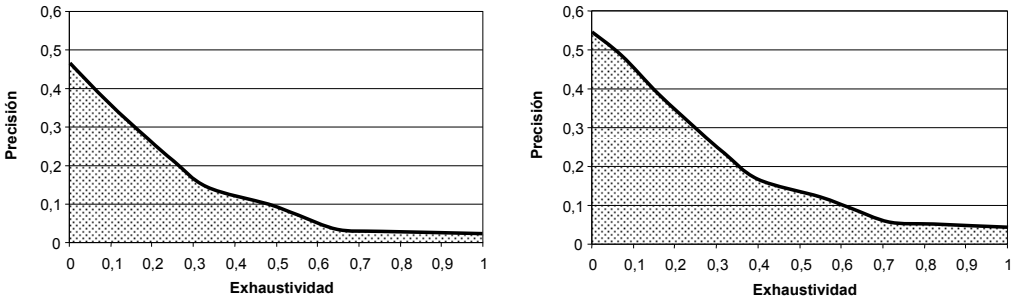
Se trata de la **precisión** de un sistema de **recuperación de información** cuando se han retornado exactamente *k* resultados. Por ejemplo, supongamos que un sistema *IR* retorna para una consulta 10 documentos de los cuales 4 son relevantes; entonces la precisión en 10 es de 0,4. Si se retornan 10 documentos más y ninguno es relevante entonces la precisión en 20 será de 0,2. Dicho de otro modo:

$$precisionEnK = \frac{\sum_{i=1}^{K} rel(i)}{K}$$

Donde *K* es el número de resultados, *i* es el resultado *i*-ésimo y *rel(i)* retorna 1 si el resultado *i*-ésimo es relevante para la consulta y 0 en caso contrario.

Precisión-Exhaustividad, curva de

Representación gráfica del comportamiento de un sistema de **recuperación de información** mostrando cómo varía la **precisión interpolada** (eje de ordenadas) con la **exhaustividad** (eje de abscisas). Una curva precisión-exhaustividad típica es cóncava y decreciente. Este tipo de representación permite comparar con relativa facilidad dos sistemas *IR* puesto que un mejor rendimiento (mayor precisión y exhaustividad) supone una mayor superficie encerrada bajo la curva.



Precisión interpolada

La **precisión en *k*** y la **precisión media** se calculan para una única consulta; sin embargo, resulta mucho más interesante obtener una medida que involucre varias consultas. Para ello es necesario (1) establecer una serie de valores estándar para la exhaustividad (típicamente 0, 0.1, 0.2, …, 0.9 y 1.0), (2) transformar los valores de precisión para cada consulta a estos puntos estandarizados y (3) calcular el valor medio para todas las consultas en cada uno de los 11 puntos. Una vez obtenidos estos valores se pueden representar mediante una **curva de precisión-exhaustividad**.

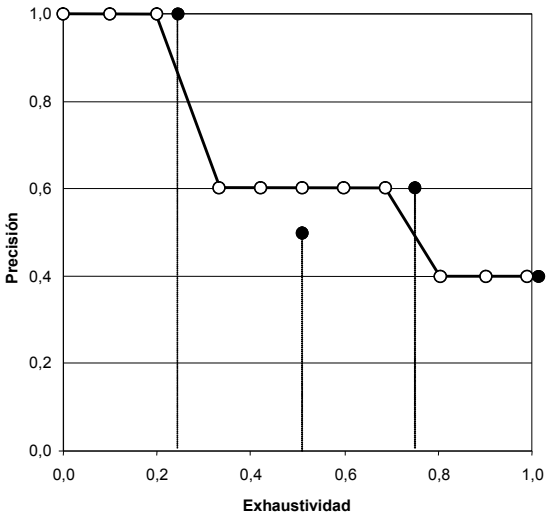
La precisión interpolada para un valor estándar de exhaustividad ρ no es más que el mayor valor de precisión para cualquier valor de exhaustividad experimental mayor o igual que ρ. Por ejemplo, dada la siguiente colección de documentos:

{A, B, C, D, E, F, G, H, I, J, K, L, M, N, O}

Supongamos que para una consulta dada la lista de resultados es la siguiente:

{**E**, L, M, **I**, **O**, C, N, D, F, **A**, H, G, B, J, K}

Donde se muestran subrayados los documentos relevantes. Así pues, se tendría que para una exhaustividad de 0.25 la precisión es de 1; para 0.5 es 0.5; para 0.75 es 0.6 y para 1 es 0.4. Las precisiones interpoladas para los valores estándar de exhaustividad serán entonces: 1 para 0, 0.1 y 0.2; 0.6 para 0.3, 0.4, 0.5, 0.6 y 0.7 y 0.4 para 0.8, 0.9 y 1. En la figura se muestran los valores de precisión obtenidos experimentalmente como círculos negros y los valores interpolados como círculos blancos; también se muestra la curva de precisión-exhaustividad.



Precisión media (no interpolada)

La precisión media es una medida única del rendimiento de un sistema de **recuperación de información** que combina **precisión**, **exhaustividad** y la calidad de la ordenación de los resultados. Se trata de la media de los distintos valores de precisión para cada uno de los documento relevantes de la colección.

Por ejemplo, dada la siguiente colección de documentos:

{A, B, C, D, E, F, G, H, I, J, K, L, M, N, O}

Supongamos que para una consulta dada la lista de resultados es la siguiente:

{**E**, L, M, **I**, **O**, C, N, D, F, **A**, H, G, B, J, K}

Donde se muestran subrayados los documentos relevantes. Para calcular la precisión media es necesario calcular la precisión cada vez que se recupera un documento relevante hasta haber completado todos los documentos relevantes. Así, los valores de precisión serían: 1 al recuperar el documento E; 0,5 al recuperar el documento I; 0,6 al recuperar el documento O y 0,4 al recuperar el documento A. Por tanto la precisión media sería:

$$\frac{1+0{,}5+0{,}6+0{,}4}{4} = \frac{2{,}5}{4} = 0{,}625$$

Dicho de otro modo:

$$precisionMedia = \frac{\sum_{r=1}^{N} P(r) \cdot rel(r)}{R}$$

Donde N es el número de resultados; r es el resultado r-ésimo; $P(r)$ es la **precisión en *r***; $rel(r)$ es 1 si el resultado r-ésimo es relevante para la consulta y 0 en caso contrario y R es el número de documentos relevantes para la consulta.

Procesamiento de Lenguaje Natural (PLN)

El Procesamiento de Lenguaje Natural (PLN) es el conjunto de técnicas algorítmicas que tienen como objeto la manipulación y generación de muestras de lenguaje humano tanto en su manifestación escrita como oral. Ejemplos de técnicas de PLN son la generación de habla a partir de texto, el reconocimiento del habla, la traducción automática o la **recuperación de información**.

Programación genética

La programación genética es una técnica para la generación automática de programas de ordenador que alcancen el mejor rendimiento posible en una tarea definida por el usuario: generalmente aproximar una función desconocida pero para la cual se conoce su comportamiento deseado para un conjunto de datos de entrada. En su forma más sencilla los programas son simples expresiones representadas en forma de árbol.

Pureza

La pureza es una medida para la evaluación de soluciones de **agrupamiento** (que producen grupos de elementos) mediante la comparación con una clasificación previa (que proporciona una serie de clases). No es más que la proporción entre el número de *ítems* pertenecientes a la clase dominante en un grupo y el tamaño de dicho grupo. Es decir, la pureza evalúa en qué medida un grupo de una clasificación automática contiene elementos de una única clase.

Recall

Véase **exhaustividad**.

Recomendación por contenidos

La recomendación por contenidos es una técnica que permite a un sistema proporcionar documentos similares a un documento de partida y que precisa, por tanto, de algún tipo de medida de **similitud** entre documentos.

Recuperación de información (*IR* o *information retrieval*)

El término recuperación de información hace referencia, en general, al estudio de sistemas automáticos que permitan a un usuario determinar la existencia o inexistencia de documentos (esto es, textos) relativos a una necesidad de información formulada habitualmente como una **consulta**.

Red neuronal

Técnica de **categorización** consistente en una estructura de capas interconectadas y formadas por elementos de procesamiento cuya funcionalidad está inspirada en las neuronas animales. Las redes neuronales requieren al menos dos capas: una de entrada con tantos elementos como variables definan a los objetos del problema y otra de salida con tantos elementos como categorías deba reconocer la red neuronal. Opcionalmente puede haber una o más capas ocultas. El aprendizaje de la red se produce mediante un entrenamiento durante el cual se ajustan los pesos de los distintos nodos de la red.

Reglas de decisión

Una técnica de categorización, no necesariamente automática, similar en cierta medida a los **árboles de decisión** y que consiste en la utilización de un conjunto de reglas para categorizar algún tipo de objetos en función de una serie de variables que los definen. Por su propia naturaleza las reglas de decisión son susceptibles de ser producidas de manera manual por los propios usuarios (p.ej. para dirigir correo electrónico con determinadas palabras en el asunto a una carpeta en particular).

Relevancia

La relevancia es una medida de la **similitud** entre los contenidos de un documento y la **consulta** de un usuario. Se trata de un valor subjetivo y cambiante, por lo que el término no suele hacer referencia al "juicio" que emitiría un usuario sino al valor que un sistema de **recuperación de información** asigna a cada documento en relación con una consulta. El objetivo de tales sistemas es producir valores de relevancia próximos a los que asignaría el propio usuario.

Resumen abstractivo/extractivo (*Abstract vs. Extract*)

Según Hovy (1999) un resumen extractivo (*extract*) consiste en una selección de parte del material presente en un documento original mientras que un resumen abstractivo (*abstract*) consiste en una condensación y reformulación del original. Un resumen extractivo se considera, a su vez, **informativo** mientras que uno abstractivo suele ser **indicativo**. Por lo que respecta al **resumen automático** resulta mucho más sencillo producir resúmenes por extracción que por abstracción.

Resumen automático (*automatic summarization*)

Las técnicas de resumen automático tienen como misión obtener a partir de un documento o conjunto de documentos un único texto mucho más corto que aún contenga los aspectos más relevantes de los originales. Durante los años 1950 y 1960 la investigación en este tipo de tecnologías fue intensa para descender considerablemente durante los años siguientes y no recuperarse hasta los años 1990. Desde entonces se trata de un campo muy activo y aunque aún se está lejos de disponer de sistemas capaces de emular a un ser humano (p.ej. produciendo

resúmenes indicativos) se ha avanzado enormemente y los sistemas estadísticos y puramente **extractivos** (o de "cortar-y-pegar") han demostrado su gran utilidad.

Resumen indicativo/informativo

Según Hovy (1999) *"un resumen informativo refleja el contenido del texto original, probablemente detallando sus argumentos, mientras que un resumen indicativo simplemente proporciona una indicación sobre el tema que trataba el documento original"*. Así, los resúmenes informativos suelen reemplazar a los documentos que resumen mientras que los indicativos permiten a los usuarios decidir sobre la pertinencia del documento en relación a una necesidad de información específica.

Ro (P)

En el contexto de la técnica ***blindLight*** se define P como el cociente de la **significatividad total** del vector **intersección** de una **consulta** y un documento entre la significatividad total de un vector documento. P revela en qué medida el documento "satisface" a la intersección entre éste y una consulta. Esta medida puede combinarse con **Pi (Π)** para construir distintas medidas de **similitud** adaptadas a las necesidades de las diferentes aplicaciones.

Rocchio, algoritmo de

Algoritmo para la expansión de **consultas** por realimentación (*relevance feedback*) que también se ha empleado como técnica de **categorización automática**. La idea subyacente a la técnica es sencilla: (1) dada una consulta, un sistema de recuperación de información proporciona al usuario un conjunto de documentos, (2) el usuario selecciona los que considera relevantes y (3) se "enriquece" la consulta original calculando la diferencia entre los documentos relevantes (POS_i, véase la ecuación) y los no relevantes (NEG_i). Así, una categoría c_i estaría representada por un vector de pesos w_{ki} calculados según la siguiente fórmula en la que w_{kj} es el peso del término t_k en el documento d_j.

$$w_{ki} = \beta \cdot \sum_{\{d_j \in POS_i\}} \frac{w_{kj}}{|POS_i|} - \gamma \cdot \sum_{\{d_j \in NEG_i\}} \frac{w_{kj}}{|NEG_i|}$$

Self-Organizing Maps (SOM)

Los Mapas Auto-Organizativos (Kohonen 1982) consisten en una **red neuronal** (generalmente bi o tridimensional) que se entrena con vectores de características en un proceso competitivo. Para cada vector hay una única neurona ganadora que ajustará sus pesos para aproximarse al vector de entrada. No obstante, el resto de neuronas también ajustan parcialmente sus pesos de forma inversamente proporcional a la distancia a que se encuentren de la vencedora. De este modo, se van vinculando los vectores a diferentes coordenadas del mapa y en caso de que estén etiquetados se asociarán sus etiquetas a las distintas zonas del mismo.

Significatividad

En el contexto de la técnica ***blindLight*** se denomina significatividad a los pesos asignados a cada uno de los *n*-gramas que forman un vector documental. Dichos pesos son calculados a partir tan sólo del texto original empleando alguno de los estadísticos propuestos por Ferreira da Silva y Pereira Lopes (1999) como la información mutua o la probabilidad condicional simétrica.

Significatividad total

En el contexto de la técnica ***blindLight*** la significatividad total es la suma de los valores de **significatividad** para todos los *n*-gramas que componen un documento.

Similitud

La similitud es una cantidad que refleja la fuerza de la asociación (o parecido) entre dos objetos en función de sus características. Suele variar en el rango [-1, 1] aunque puede normalizarse entre 0 y 1. La similitud normalizada puede transformarse de manera trivial en una **disimilitud** normalizada, si s_{ij} es la similitud normalizada entre los elementos *i* y *j* entonces la disimilitud entre ambos será:

$$\delta_{ij} = 1 - s_{ij}$$

En el **modelo vectorial** se emplea habitualmente la **función del coseno** para calcular la similitud entre documentos.

Similitud promedio (*overall similarity*)

La similitud promedio es una medida para la evaluación de soluciones de **agrupamiento** (que producen grupos de elementos) que no precisa de la comparación con una clasificación previa. Se trata tan sólo de la **similitud** media entre cada par de documentos de un grupo.

Stemming (reducción a la raíz)

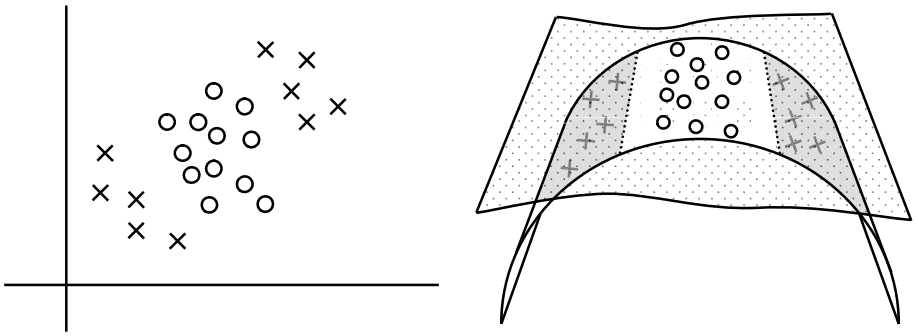
Un algoritmo de *stemming* o *stemmer* determina la raíz morfológica de una palabra colapsando múltiples formas de la raíz en un único término (`research`, `researcher`, `researching` y `researchers` colapsan en `research` empleando un *stemmer* para inglés). Un *stemmer* para castellano, por ejemplo, transformaría `andanzas` en `and`, `habitaciones` en `habit` o `jugueis` en `jug`.

Stop words

Se denominan *stop words* o palabras vacías aquellas palabras que, a pesar de un uso frecuente, aportan por sí solas poco significado a un texto. En la sentencia anterior se muestran subrayadas algunas palabras vacías del castellano.

Support Vector Machines (*SVM*)

Método de **categorización automática** propuesto por Boser, Guyon y Vapnik (1992) y que consiste en la transformación de los vectores de entrada (véase **modelo vectorial**), que definen una dimensión en la cual no son linealmente separables, a una dimensión superior que permita su separación mediante una única (hiper)superficie. Se trata de una de las técnicas de categorización más eficientes aunque, tal vez debido a su complejidad, otras que ofrecen peores resultados (como los **categorizadores bayesianos**) continuan siendo muy populares.



*tf*idf*

Método para ponderar los términos de un documento en un sistema de **recuperación de información**. Según este método la importancia de un término es directamente proporcional a su frecuencia de aparición en un documento (*tf*) e inversamente proporcional al número de documentos en que aparece (***idf***).

Valoración explícita/implícita

En los sistemas de **filtrado colaborativo** se recomiendan nuevos elementos a un usuario en función de sus preferencias pasadas y de la utilidad que dichos elementos para otros usuarios con preferencias similares. Para determinar dicha utilidad es necesaria una valoración por parte del usuario; dicha valoración puede ser explícita (p.ej. puntuando el elemento) o implícita, es decir, sin requerir la intervención del usuario y basándose tan sólo en su interacción con el elemento (p.ej. tiempo de lectura, impresión del documento, incorporación a la lista de enlaces favoritos, etc.)

Web Semántica

Según Tim Berners-Lee, James Hendler y Ora Lassila (2001) *"la Web Semántica es una extensión de la Web actual en la cual se asigna a la información un significado bien definido, posibilitando una mejor cooperación entre máquinas y usuarios"*. Una pieza clave en el desarrollo de la Web Semántica son las **ontologías**.

ANEXO: MÉTODO *BLINDLIGHT* PARA RESÚMEN AUTOMÁTICO

ROUGE-1	Método de resumen muy corto (75 caracteres)	Dif. Top 5	Dif. Media	Percentil
0,29236	Humano			
0,25033	Mejor participante DUC 2004			
0,22553	Mejores 5 participantes DUC 2004			
0,22136	Baseline			
0,19081	**Dice, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave**	**-15,4%**	**3,8%**	**59**
0,18385	Media DUC 2004			
0,17690	Dice, 3-gramas, ventana 6, palabras clave únicas			
0,17455	SCP, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,16701	mutual information, 3-gramas, ventana 6, palabras clave ordenadas			
0,16552	SCP, 3-gramas, ventana 6, palabras clave únicas			
0,16048	mutual information, 3-gramas, ventana 8, construido a partir de fragmentos por puntuación por palabras clave			
0,16018	mutual information, 3-gramas, ventana 7, construido a partir de fragmentos por puntuación por palabras clave			
0,15789	mutual information, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,15728	mutual information, 3-gramas, ventana 5, construido a partir de fragmentos por puntuación por palabras clave			
0,15646	Dice, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,15615	mutual information, 3-gramas, ventana 6, palabras clave únicas			
0,14498	mutual information, 3-gramas, ventana 5%, construido a partir de fragmentos por puntuación por palabras clave			
0,14472	mutual information, 3-gramas, ventana 10, construido a partir de fragmentos por puntuación por palabras clave			
0,14448	mutual information, 3-gramas, ventana 10%, construido a partir de fragmentos por puntuación por palabras clave			
0,14289	Dice, 3-gramas, sentencia más significativa, sin "compresión"			
0,14213	mutual information, 3-gramas, ventana 3, construido a partir de fragmentos por puntuación por palabras clave			
0,14122	SCP, 3-gramas, sentencia más significativa, sin "compresión"			
0,14047	mutual information, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,14003	SCP, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,13623	Infogain, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,13527	mutual information, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,13317	Dice, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,13245	Infogain, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,13237	Dice, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,13218	mutual information, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,13119	mutual information, 3-gramas, sentencia más significativa, sin "compresión"			
0,13081	SCP, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,12710	mutual information, 4-gramas, sentencia más significativa, sin "compresión"			
0,12702	Infogain, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,12523	SCP, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,12518	Infogain, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,11834	mutual information, 3-gramas, sentencia mayor puntuación por fragmentos, con "compresión"			
0,11733	Chi2, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,11691	Chi2, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,11589	Chi2, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,11502	mutual information, 5-gramas, sentencia más significativa, sin "compresión"			
0,11501	mutual information, 5-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,11403	Infogain, 3-gramas, sentencia más significativa, sin "compresión"			
0,11246	mutual information, 4-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,11163	Chi2, 3-gramas, ventana 6, palabras clave únicas			
0,11057	mutual information, 4-gramas, sentencia más significativa, con "compresión"			
0,10948	mutual information, 4-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,10823	Infogain, 3-gramas, ventana 6, palabras clave únicas			
0,10619	Chi2, 3-gramas, sentencia más significativa, sin "compresión"			
0,10201	Chi2, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,09470	mutual information, 5-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			

ROUGE-2	Método de resumen muy corto (75 caracteres)	Dif. Top 5	Dif. Media	Percentil
0,08411	Humano			
0,06501	Mejor participante DUC 2004			
0,06370	Baseline			
0,06190	Mejores 5 participantes DUC 2004			
0,03973	Media DUC 2004			
0,02964	**mutual information, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"**	**-52,1%**	**-25,4%**	**24**
0,02853	**Dice, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"**			
0,02627	mutual information, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,02559	SCP, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,02518	mutual information, 3-gramas, ventana 10, construido a partir de fragmentos por puntuación por palabras clave			
0,02500	mutual information, 3-gramas, ventana 10%, construido a partir de fragmentos por puntuación por palabras clave			
0,02486	mutual information, 3-gramas, ventana 5, construido a partir de fragmentos por puntuación por palabras clave			
0,02464	mutual information, 3-gramas, ventana 7, construido a partir de fragmentos por puntuación por palabras clave			
0,02460	mutual information, 3-gramas, ventana 8, construido a partir de fragmentos por puntuación por palabras clave			
0,02429	mutual information, 3-gramas, ventana 5%, construido a partir de fragmentos por puntuación por palabras clave			
0,02422	mutual information, 4-gramas, sentencia más significativa, sin "compresión"			
0,02422	Infogain, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,02390	mutual information, 3-gramas, ventana 3, construido a partir de fragmentos por puntuación por palabras clave			
0,02311	mutual information, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,02217	Dice, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,02188	mutual information, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,02162	Infogain, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,02131	mutual information, 3-gramas, sentencia más significativa, sin "compresión"			
0,02062	SCP, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,01998	mutual information, 5-gramas, sentencia más significativa, sin "compresión"			
0,01929	Dice, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,01929	Dice, 3-gramas, sentencia más significativa, sin "compresión"			
0,01831	SCP, 3-gramas, sentencia más significativa, sin "compresión"			
0,01827	mutual information, 4-gramas, sentencia más significativa, con "compresión"			
0,01813	Chi2, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,01793	mutual information, 4-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,01773	SCP, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,01740	Chi2, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,01613	Chi2, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,01568	mutual information, 5-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,01473	mutual information, 4-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,01428	mutual information, 3-gramas, sentencia mayor puntuación por fragmentos, con "compresión"			
0,01384	Chi2, 3-gramas, sentencia más significativa, sin "compresión"			
0,01224	Infogain, 3-gramas, sentencia más significativa, sin "compresión"			
0,01170	mutual information, 3-gramas, ventana 6, palabras clave ordenadas			
0,01165	Infogain, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,01112	mutual information, 5-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,01026	mutual information, 3-gramas, ventana 6, palabras clave únicas			
0,00732	Dice, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,00614	Infogain, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,00565	Dice, 3-gramas, ventana 6, palabras clave únicas			
0,00509	SCP, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,00446	SCP, 3-gramas, ventana 6, palabras clave únicas			
0,00428	Infogain, 3-gramas, ventana 6, palabras clave únicas			
0,00211	Chi2, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,00184	Chi2, 3-gramas, ventana 6, palabras clave únicas			

ROUGE-3	Método de resumen muy corto (75 caracteres)	Dif. Top 5	Dif. Media	Percentil
0,02897	Humano			
0,02134	Mejor participante DUC 2004			
0,02118	Baseline			
0,02025	Mejores 5 participantes DUC 2004			
0,01065	Media DUC 2004			
0,01041	**mutual information, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"**	**-48,6%**	**-2,3%**	**56**
0,00784	mutual information, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,00783	Infogain, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,00773	mutual information, 4-gramas, sentencia más significativa, sin "compresión"			
0,00722	Dice, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,00705	SCP, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,00700	mutual information, 3-gramas, ventana 7, construido a partir de fragmentos por puntuación por palabras clave			
0,00681	mutual information, 3-gramas, ventana 8, construido a partir de fragmentos por puntuación por palabras clave			
0,00644	mutual information, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,00640	mutual information, 3-gramas, ventana 10, construido a partir de fragmentos por puntuación por palabras clave			
0,00637	mutual information, 3-gramas, ventana 5%, construido a partir de fragmentos por puntuación por palabras clave			
0,00636	mutual information, 3-gramas, ventana 5, construido a partir de fragmentos por puntuación por palabras clave			
0,00486	mutual information, 3-gramas, ventana 10%, construido a partir de fragmentos por puntuación por palabras clave			
0,00428	Dice, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,00408	Infogain, 3-gramas, sentencia más significativa, sin "compresión"			
0,00405	Dice, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,00396	SCP, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,00393	Chi2, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,00387	mutual information, 5-gramas, sentencia más significativa, sin "compresión"			
0,00383	Infogain, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,00377	mutual information, 3-gramas, sentencia más significativa, sin "compresión"			
0,00375	SCP, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,00369	mutual information, 3-gramas, ventana 3, construido a partir de fragmentos por puntuación por palabras clave			
0,00365	mutual information, 4-gramas, sentencia más significativa, con "compresión"			
0,00360	Chi2, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,00359	mutual information, 4-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,00358	Chi2, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,00347	Dice, 3-gramas, sentencia más significativa, sin "compresión"			
0,00343	Chi2, 3-gramas, sentencia más significativa, sin "compresión"			
0,00337	mutual information, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,00296	mutual information, 3-gramas, sentencia mayor puntuación por fragmentos, con "compresión"			
0,00260	Infogain, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,00227	mutual information, 5-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,00116	mutual information, 5-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,00115	SCP, 3-gramas, sentencia más significativa, sin "compresión"			
0,00112	mutual information, 4-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,00031	Infogain, 3-gramas, ventana 6, palabras clave únicas			
0,00031	Infogain, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,00021	Dice, 3-gramas, ventana 6, palabras clave únicas			
0,00021	Dice, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,00020	mutual information, 3-gramas, ventana 6, palabras clave únicas			
0,00020	mutual information, 3-gramas, ventana 6, palabras clave ordenadas			
0,00017	SCP, 3-gramas, ventana 6, palabras clave únicas			
0,00017	SCP, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,00014	Chi2, 3-gramas, ventana 6, palabras clave únicas			
0,00014	Chi2, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			

ROUGE-4	Método de resumen muy corto (75 caracteres)	Dif. Top 5	Dif. Media	Percentil
0,01051	Humano			
0,00730	Mejor participante DUC 2004			
0,00707	Baseline			
0,00679	Mejores 5 participantes DUC 2004			
0,00298	Media DUC 2004			
0,00290	**mutual information, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"**	**-57,3%**	**-2,7%**	**57**
0,00274	mutual information, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,00274	mutual information, 3-gramas, top fragmentosig, sin "compresión"			
0,00268	mutual information, 3-gramas, ventana 7, construido a partir de fragmentos por puntuación por palabras clave			
0,00253	Dice, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,00247	SCP, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,00220	Infogain, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,00210	mutual information, 3-gramas, ventana 8, construido a partir de fragmentos por puntuación por palabras clave			
0,00210	Infogain, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,00210	Dice, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,00207	mutual information, 3-gramas, ventana 10, construido a partir de fragmentos por puntuación por palabras clave			
0,00203	SCP, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,00203	SCP, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,00203	mutual information, 5-gramas, sentencia más significativa, sin "compresión"			
0,00203	mutual information, 3-gramas, ventana 5, construido a partir de fragmentos por puntuación por palabras clave			
0,00199	mutual information, 3-gramas, ventana 5%, construido a partir de fragmentos por puntuación por palabras clave			
0,00195	mutual information, 4-gramas, sentencia más significativa, sin "compresión"			
0,00195	Chi2, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,00191	Chi2, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,00187	mutual information, 3-gramas, ventana 3, construido a partir de fragmentos por puntuación por palabras clave			
0,00163	mutual information, 4-gramas, top fragmentosig, sin "compresión"			
0,00160	mutual information, 5-gramas, top fragmentosig, sin "compresión"			
0,00159	mutual information, 3-gramas, ventana 10%, construido a partir de fragmentos por puntuación por palabras clave			
0,00155	Chi2, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,00146	Infogain, 3-gramas, sentencia más significativa, sin "compresión"			
0,00124	mutual information, 3-gramas, sentencia más significativa, sin "compresión"			
0,00123	mutual information, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,00121	Chi2, 3-gramas, sentencia más significativa, sin "compresión"			
0,00119	Dice, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,00119	Dice, 3-gramas, sentencia más significativa, sin "compresión"			
0,00114	SCP, 3-gramas, sentencia más significativa, sin "compresión"			
0,00080	mutual information, 4-gramas, sentencia más significativa, con "compresión"			
0,00076	mutual information, 4-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,00068	mutual information, 5-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,00065	mutual information, 3-gramas, top fragmentosig, con "compresión"			
0,00055	Infogain, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,00007	Dice, 3-gramas, ventana 6, palabras clave únicas			
0,00007	Dice, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,00003	SCP, 3-gramas, ventana 6, palabras clave únicas			
0,00003	SCP, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,00003	mutual information, 3-gramas, ventana 6, palabras clave únicas			
0,00003	mutual information, 3-gramas, ventana 6, palabras clave ordenadas			
0,00003	Infogain, 3-gramas, ventana 6, palabras clave únicas			
0,00003	Infogain, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,00003	Chi2, 3-gramas, ventana 6, palabras clave únicas			
0,00003	Chi2, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			

ROUGE-L	Método de resumen muy corto (75 caracteres)	Dif. Top 5	Dif. Media	Percentil
0,24588	Humano			
0,20074	Mejor participante DUC 2004			
0,19411	Baseline			
0,19334	Mejores 5 participantes DUC 2004			
0,15635	Media DUC 2004			
0,11454	**mutual information, 3-gramas, ventana 7, construido a partir de fragmentos por puntuación por palabras clave**	**-40,8%**	**-26,7%**	**7**
0,11427	**mutual information, 3-gramas, ventana 8, construido a partir de fragmentos por puntuación por palabras clave**			
0,11367	**mutual information, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"**			
0,11081	**mutual information, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave**			
0,10922	**mutual information, 3-gramas, ventana 5, construido a partir de fragmentos por puntuación por palabras clave**			
0,10908	Dice, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,10759	mutual information, 3-gramas, ventana 10, construido a partir de fragmentos por puntuación por palabras clave			
0,10626	mutual information, 3-gramas, ventana 10%, construido a partir de fragmentos por puntuación por palabras clave			
0,10613	mutual information, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,10606	mutual information, 3-gramas, ventana 5%, construido a partir de fragmentos por puntuación por palabras clave			
0,10499	SCP, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,10467	mutual information, 3-gramas, ventana 3, construido a partir de fragmentos por puntuación por palabras clave			
0,10079	mutual information, 3-gramas, sentencia mayor puntuación por fragmentos, con "compresión"			
0,10047	Infogain, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,09932	mutual information, 3-gramas, ventana 6, palabras clave únicas			
0,09880	Dice, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,09863	mutual information, 4-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,09853	mutual information, 4-gramas, sentencia más significativa, sin "compresión"			
0,09821	mutual information, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,09821	Dice, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,09770	mutual information, 3-gramas, ventana 6, palabras clave ordenadas			
0,09686	Infogain, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,09632	SCP, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,09617	SCP, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,09542	Dice, 3-gramas, ventana 6, palabras clave únicas			
0,09523	mutual information, 4-gramas, sentencia más significativa, con "compresión"			
0,09491	mutual information, 5-gramas, sentencia más significativa, sin "compresión"			
0,09415	mutual information, 3-gramas, sentencia más significativa, sin "compresión"			
0,09338	Dice, 3-gramas, sentencia más significativa, sin "compresión"			
0,09288	Chi2, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,09272	SCP, 3-gramas, sentencia más significativa, sin "compresión"			
0,09181	Dice, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,09102	Chi2, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,09084	Infogain, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,09079	SCP, 3-gramas, ventana 6, palabras clave únicas			
0,08885	mutual information, 5-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,08852	SCP, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,08822	mutual information, 5-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,08759	mutual information, 4-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,08494	Chi2, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,08250	Infogain, 3-gramas, sentencia más significativa, sin "compresión"			
0,07956	Chi2, 3-gramas, sentencia más significativa, sin "compresión"			
0,07461	Chi2, 3-gramas, ventana 6, palabras clave únicas			
0,07433	Chi2, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,07427	Infogain, 3-gramas, ventana 6, palabras clave únicas			
0,06554	Infogain, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			

ROUGE-W-1.2	Método de resumen muy corto (75 caracteres)	Dif.Top 5	Dif. Media	Percentil
0,14124	Humano			
0,11957	Mejor participante DUC 2004			
0,11738	Baseline			
0,11583	Mejores 5 participantes DUC 2004			
0,09366	Media DUC 2004			
0,07078	**mutual information, 3-gramas, ventana 7, construido a partir de fragmentos por puntuación por palabras clave**	**-38,9%**	**-24,4%**	**7**
0,07063	**mutual information, 3-gramas, ventana 8, construido a partir de fragmentos por puntuación por palabras clave**			
0,07031	**mutual information, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"**			
0,06875	**mutual information, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave**			
0,06789	**Dice, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"**			
0,06781	**mutual information, 3-gramas, ventana 5, construido a partir de fragmentos por puntuación por palabras clave**			
0,06700	mutual information, 3-gramas, ventana 10, construido a partir de fragmentos por puntuación por palabras clave			
0,06628	mutual information, 3-gramas, ventana 10%, construido a partir de fragmentos por puntuación por palabras clave			
0,06621	mutual information, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,06617	mutual information, 3-gramas, ventana 5%, construido a partir de fragmentos por puntuación por palabras clave			
0,06559	SCP, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,06542	mutual information, 3-gramas, ventana 3, construido a partir de fragmentos por puntuación por palabras clave			
0,06331	mutual information, 3-gramas, sentencia mayor puntuación por fragmentos, con "compresión"			
0,06313	Infogain, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,06251	mutual information, 3-gramas, ventana 6, palabras clave únicas			
0,06223	Dice, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,06213	mutual information, 4-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,06208	mutual information, 4-gramas, sentencia más significativa, sin "compresión"			
0,06191	Dice, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,06190	mutual information, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,06163	mutual information, 3-gramas, ventana 6, palabras clave ordenadas			
0,06117	Infogain, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,06088	SCP, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,06080	SCP, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,06039	Dice, 3-gramas, ventana 6, palabras clave únicas			
0,06029	mutual information, 4-gramas, sentencia más significativa, con "compresión"			
0,06011	mutual information, 5-gramas, sentencia más significativa, sin "compresión"			
0,05970	mutual information, 3-gramas, sentencia más significativa, sin "compresión"			
0,05928	Dice, 3-gramas, sentencia más significativa, sin "compresión"			
0,05901	Chi2, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,05892	SCP, 3-gramas, sentencia más significativa, sin "compresión"			
0,05843	Dice, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,05800	Chi2, 3-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,05790	SCP, 3-gramas, ventana 6, palabras clave únicas			
0,05787	Infogain, 3-gramas, ventana 6, construido a partir de fragmentos con mayor puntuación por palabras clave			
0,05682	mutual information, 5-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,05664	SCP, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,05648	mutual information, 5-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, con "compresión"			
0,05613	mutual information, 4-gramas, sentencia mayor puntuación por fragmentos, sin "compresión"			
0,05469	Chi2, 3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, sin "compresión"			
0,05337	Infogain, 3-gramas, sentencia más significativa, sin "compresión"			
0,05177	Chi2, 3-gramas, sentencia más significativa, sin "compresión"			
0,04908	Chi2, 3-gramas, ventana 6, palabras clave únicas			
0,04893	Infogain, 3-gramas, ventana 6, palabras clave únicas			
0,04890	Chi2, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			
0,04415	Infogain, 3-gramas, ventana 6, construido a partir de fragmentos por puntuación por palabras clave			

ROUGE-1	Método de resumen corto (665 caracteres)	Dif. Top 5	Dif. Media	Percentil
0,40300	Humano			
0,38224	Mejor participante DUC 2004			
0,37509	Mejores 5 participantes DUC 2004			
0,36905	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, SCP**			
0,36902	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Dice**			
0,35496	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, mutual information**			
0,35406	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Infogain**			
0,35274	**4-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, mutual information**	**-6,0%**	**4,6%**	**65**
0,34392	3-gramas, sentencia más significativa, mutual information			
0,34078	3-gramas, sentencia mayor puntuación por fragmentos, mutual information			
0,33720	Media DUC 2004			
0,32419	Baseline			
0,32050	3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Chi2			

ROUGE-2	Método de resumen corto (665 caracteres)	Dif. Top 5	Dif. Media	Percentil
0,09592	Humano			
0,09216	Mejor participante DUC 2004			
0,08831	Mejores 5 participantes DUC 2004			
0,08490	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Dice**			
0,08354	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, SCP**			
0,08230	**4-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, mutual information**	**-6,8%**	**19,7%**	**81**
0,08205	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, mutual information**			
0,07220	3-gramas, sentencia más significativa, mutual information			
0,06873	Media DUC 2004			
0,06537	3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Infogain			
0,06411	Baseline			
0,06007	3-gramas, sentencia mayor puntuación por fragmentos, mutual information			
0,05262	3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Chi2			

ROUGE-3	Método de resumen corto (665 caracteres)	Dif. Top 5	Dif. Media	Percentil
0,03012	Humano			
0,03529	Mejor participante DUC 2004			
0,03196	Mejores 5 participantes DUC 2004			
0,02860	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Dice**			
0,02818	**4-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, mutual information**	**-11,8%**	**30%**	**87**
0,02731	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, SCP**			
0,02468	3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, mutual information			
0,02165	Media DUC 2004			
0,01992	Baseline			
0,01866	3-gramas, sentencia más significativa, mutual information			
0,01837	3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Infogain			
0,01701	3-gramas, sentencia mayor puntuación por fragmentos, mutual information			
0,01483	3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Chi2			

ROUGE-4	Método de resumen corto (665 caracteres)	Dif. Top 5	Dif. Media	Percentil
0,01658	Mejor participante DUC 2004			
0,01464	Mejores 5 participantes DUC 2004			
0,01267	**4-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, mutual information**	**-13,5%**	**41,4%**	**88**
0,01158	3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Dice			
0,01091	Humano			
0,01090	3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, SCP			
0,00943	3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, mutual information			
0,00896	Media DUC 2004			
0,00687	3-gramas, sentencia más significativa, mutual information			
0,00679	3-gramas, sentencia mayor puntuación por fragmentos, mutual information			
0,00671	Baseline			
0,00641	3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Chi2			
0,00629	3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Infogain			

ROUGE-L	Método de resumen corto (665 caracteres)	Dif. Top 5	Dif. Media	Percentil
0,42018	Humano			
0,38950	Mejor participante DUC 2004			
0,38570	Mejores 5 participantes DUC 2004			
0,36583	**4-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, mutual information**	**-5,2%**	**5,6%**	**66**
0,36289	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, SCP**			
0,36288	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Dice**			
0,36098	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Infogain**			
0,36069	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, mutual information**			
0,35530	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Chi2**			
0,34642	Media DUC 2004			
0,34590	Baseline			
0,34204	3-gramas, sentencia más significativa, mutual information			
0,30875	3-gramas, sentencia mayor puntuación por fragmentos, mutual information			

ROUGE-W-1.2	Método de resumen corto (665 caracteres)	Dif. Top 5	Dif. Media	Percentil
0,14288	Humano			
0,13378	Mejor participante DUC 2004			
0,13241	Mejores 5 participantes DUC 2004			
0,12591	**4-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, mutual information**	**-4,9%**	**5,7%**	**67**
0,12348	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, SCP**			
0,12348	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Dice**			
0,12279	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, mutual information**			
0,12183	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Infogain**			
0,12156	**3-gramas, combinación de mayor puntuación por fragmentos y mayor significatividad por carácter, Chi2**			
0,11913	Media DUC 2004			
0,11859	Baseline			
0,11821	3-gramas, sentencia más significativa, mutual information			
0,10588	3-gramas, sentencia mayor puntuación por fragmentos, mutual information			

REFERENCIAS

Abir, E. 2003a, *Content conversion method and apparatus*, US Pat. App. 20030061025.

Abir, E. 2003b, *Multilingual database creation system and method*, US Pat. App. 20030139920.

Abir, E. 2003c, *Word association method and apparatus*, US Pat. App. 20030171910.

Abir, E. 2004, *Knowledge system method and appparatus*, US Pat. App. 20040122656.

Abir, E., Klein, S., Miller, D. y Steinbaum, M. 2002, "Fluent Machines' EliMT System", en S.D. Richardson (Ed.): *AMTA 2002*, LNAI 2499, pp. 216-219.

Aduna B.V y Sirma AI Ltd. 2004, "The SeRQL query language", en *User Guide for Sesame*, [Online], openRDF.org, Disponible en: <http://www.openrdf.org/doc/users/ch06.html> [30 Junio 2005].

Aguillo, I.F. 2002, *Measuring informal scientific publication in the Web*, [Online], EICSTES, Disponible en: <http://www.eicstes.org/EICSTES_PDF/PRESENTATIONS/Measuring%20informal%20scientific%20publication%20in%20the%20Web%20(Aguillo).PDF> [30 Junio 2005].

Alfonseca, E. y Rodríguez, P. "Description of the UAM system for generating very short summaries at DUC-2003", *HLT-NAACL Text Summarization Workshop and Document Understanding Conference 2003*.

Alfonseca, E., Guirao, J.M. y Moreno Sandoval, A. 2004, "Description of the UAM system for generating very short summaries at DUC-2004", *NAACL Text Summarization Workshop and Document Understanding Conference 2004*.

Allan, J. 1995, "Relevance Feedback With Too Much Data", en *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.*

Androutsopoulos, I., Koutsias, J., Chandrinos, K.V., Paliouras, G. y Spyropoulos, C.D. 2000a, "An Evaluation of Naive Bayesian Anti-Spam Filtering", [Online], en Potamias, G., Moustakis, V. y van Someren, M. (Eds.): *Proceedings of the Workshop on Machine Learning in the New Information Age*, pp. 9-17, disponible en: <http://www.aueb.gr/users/ion/docs/mlnet_paper.pdf> [30 Junio 2005]

Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C.D. y Stamatopoulos, P. 2000b, "Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach", [Online], en Zaragoza, H., Gallinari, P. y Rajman, M. (Eds.): *Proceedings of the Workshop on Machine Learning and Textual Information Access*, pp. 1-13, disponible en: <http://arxiv.org/abs/cs/0009009> [30 Junio 2005]

Apté, C., Damerau, F. y Weiss, S.M. 1994, "Automated Learning of Decision Rules for Text Categorization", *ACM Transactions on Information Systems*, vol. 12, no. 3, pp. 233-251.

Apté, C., Damerau, F. y Weiss, S.M. 1998, "Maximizing Text-Mining Performance", *IEEE Intelligent Systems*, vol. 14, no. 4, pp. 63-69.

Attardi, G., Gullì, A. y Sebastiani, F. 1999, "Theseus: Categorization by context", en *Proceedings of WWW8*, pp. 136-137

Baclace, P.E. 1991, "Personal Information Intake Filtering", *Bellcore Information Filtering Workshop.*

Baclace, P.E. 1992, "Competitive agents for information filtering", *Communications of the ACM*, vol. 35, no. 12, p. 50.

Baeza-Yates, R. y Ribeiro-Neto, B. 1999, *Modern Information Retrieval*, ACM Press.

Balabanovic, M. 1998, "An interface for learning multi-topic user profiles from implicit feedback", en *Proceedings of AAAI Workshop on Recommender Systems*, pp. 6-10.

Balabanovic, M. y Shoham, Y. 1997, "Fab: Content-Based, Collaborative Recommendation", *Communications of the ACM*, vol. 40, no. 3, pp. 66-72.

Balabanovic, M., Shoham, Y. y Yun, Y. 1995, An adaptive agent for automated web browsing, Informe técnico, Stanford University.

Barton, I.J., Creasey, S.E., Lynch, M.F., Snell, M.J. 1974, "An Information-Theoretic Approach to Text Searching in Direct Access Systems", *Communications of the ACM*, vol. 17, no. 6, pp. 345-350.

Barzilay, R. y Elhadad, M. 1997, "Using Lexical Chains for Text Summarization", en *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, pp. 10-17.

Barzilay, R. y Lee, L. 2002, "Bootstrapping Lexical Choice via Multiple-Sequence Alignment", en *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 164-171.

Barzilay, R. y Lee, L. 2003, "Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment", en *Proceedings of HLT-NAACL 2003*, pp. 16-23.

Bates, M.J. 1989, "The Design of Browsing and Berrypicking Techniques for the Online Search Interface.", *Online Review*, 13, pp. 407-424.

Bates, M.J. 2002, "After the Dot-Bomb: Getting Web Information Retrieval Right This Time", *First Monday*, vol. 7, no. 7, [Online], Disponible en: <http://firstmonday.org/issues/issue7_7/bates/index.html> [30 Junio 2005].

Baxendale, P.B. 1958, "Machine-made index for technical literature – an experiment", *IBM Journal*, pp. 354-361.

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L, Patel Schneider, P.F. y Stein, L.A. 2004, *OWL Web Ontology Language Reference*, [Online], Recomendación W3C, Consorcio W3, Disponible en: <http://www.w3.org/TR/owl-ref/> [30 Junio 2005].

Beesley, K.R. 1988, "Language identifier: A computer program for automatic natural-language identification on on-line text", en *Proceedings of the 29th Annual Conference of the American Translators Association*, pp. 47-54.

Belkin, N. y Vickery, A. 1985, *Interaction in Information Systems: A Review of Research from Document Retrieval to Knowledge-based systems (LIR Report No 35)*, The British Library, Reino Unido.

Bennett, C.H., Gács, P., Li, M., Vitányi, M.B. y Zurek, W.H. 1998, "Thermodynamics of Computation and Information Distance", *IEEE Transactions in Information Theory*, vol. 44, no. 4, pp. 1407-1423.

Bergman, M.K. 2001, "The Deep Web: Surfacing Hidden Value", *The Journal of Electronic Publishing*, vol. 7, no. 1, Disponible en: <http://www.press.umich.edu/jep/07-01/bergman.html> [30 Junio 2005].

Berners-Lee, T. 1989, *Information Management: A Proposal*, Informe técnico, CERN.

Berners-Lee, T. 1990, *HyperText Design Issues: Topology*, [Online], W3 Consortium, Disponible en: <http://www.w3.org/DesignIssues/Topology.html> [30 Junio 2005].

Berners-Lee, T. 1992a, *Web of Indexes*, [Online], W3 Consortium, Disponible en: <http://www.w3.org/History/19921103-hypertext/hypertext/WWW/DesignIssues/ManyIndexes.html> [30 Junio 2005].

Berners-Lee, T. 1992b, *Tracing Links*, [Online], W3 Consortium, Disponible en: <http://www.w3.org/History/19921103-hypertext/hypertext/WWW/DesignIssues/TracingLinks.html> [30 Junio 2005].

Berners-Lee, T. 1992c, *World-Wide Web Servers*, [Online], W3 Consortium, Disponible en: <http://www.w3.org/History/19921103-hypertext/hypertext/DataSources/WWW/Servers.html> [30 Junio 2005].

Berners-Lee, T. 1998, *Semantic Web Road map*, [Online], Consorcio W3, Disponible en: <http://www.w3.org/DesignIssues/Semantic.html> [30 Junio 2005].

Berners-Lee, T., Hendler, J. y Lassila, O. 2001, "The Semantic Web", *Scientific American*, vol. 284, no. 5, pp. 34-43.

Bharat, K., y Henzinger, M. 1998, "Improved Algorithms for Topic Distillation in a Hyperlinked Environment", en *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pp. 104-111.

Billhardt, H., Borrajo, D. y Maojo, V. 2003, "Learning retrieval expert combinations with genetic algorithms", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 11, no. 1, pp. 87-114.

Blair-Goldensohn, S., Evans, D., Hatzivassiloglou, V., McKeown, K.R., Nenkova, A., Passonneau, R., Schiffman, B., Schlaikjer, A., Siddharthan, A. y Siegelman, S. 2004, "Columbia University at DUC 2004", en *Document Understanding Conference at HLT-NAACL*.

Bookstein, A. y Kraft, D. 1977, "Operations research applied to document indexing and retrieval decisions", *Journal of the ACM*, vol. 24, pp. 410-427 (citado por van Rijsbergen, C.J. 1979, Information Retrieval, 2nd Edition. Butterworth-Heinemann, EE.UU.)

Boser, B., Guyon, I. y Vapnik, V. 1992, "A training algorithm for optimal margin classifiers", en *Fifth Annual Workshop on Computational Learning Theory*, pp. 144-152.

Brandow, R., Mitze, K. y Rau, L.F. 1995, "Automatic condensation of electronic publications by sentence selection", *Information Processing and Management*, vol. 31, no. 5, pp. 675-685.

Bray, J.R. y Curtis, J.T. 1957, "An ordination of the upland forest communities of southern Wisconsin". *Ecological Monographs*, 27, pp. 325-349. Citado por Gauch, H.G. 1982, *Multivariate Analysis in Community Ecology*, Cambridge University Press, EE.UU.

Brickley, D. y Guha, R.V. 2004, *RDF Vocabulary Description Language 1.0: RDF Schema*, Recomendación W3C, Consorcio W3, [Online], Disponible en: <http://www.w3.org/TR/rdf-schema> [30 Junio 2005].

Brickley, D. y Miller, L. 2000, *RDF: Extending and Querying RSS channels*, [Online], Informe técnico, Universidad de Bristol, Disponible en: <http://ilrt.org/discovery/2000/11/rss-query/> [30 Junio 2005].

Brin, S. y Page, L. 1998, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117.

Broadman, E. 1944, "Choosing physiology journals", *Bulletin of the Medical Library Association*, vol. 32, no. 4, pp. 479-483.

Broder, A. 2002, "A taxonomy of web search", *ACM SIGIR Forum*, vol. 36, no. 2, pp. 3-10.

Bruce, T.R. 1993, *Resource discovery and the Web*, [Online], University of Calgary, Disponible en: <http://ksi.cpsc.ucalgary.ca/archives/WWW-TALK/www-talk-1993q2.messages/36.html> [30 Junio 2005].

Brunn, M., Chali, Y. y Pinchak, C.J. 2001, "Text Summarization Using Lexical Chains", en *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Buckley, C., Salton, G., Allan, J. y Singhal, A. 1994, "Automatic Query Expansion Using SMART: TREC-3", en *Text REtrieval Conference*, pp. 69-80.

Burke, R. 1999, "Integrating Knowledge-based and Collaborative-filtering Recommender Systems", en *Proceedings of the AAAI Workshop on AI and Electronic Commerce*, pp. 69-72.

Busemann, S. 2001, "Language Generation for Cross-Lingual Document Summarisation", en *Proceedings of the International Workshop on Innovative Language Technology and Chinese Information Processing (ILT&CIP'01).*

Cancedda, N., Déjean, H., Gaussier, É., Renders, J.M. y Vinokourov, A. 2003, "Report on CLEF-2003 Experiments: Two Ways of Extracting Multilingual Resources from Corpora", *CLEF 2003*, pp. 98-107.

Carpineto, C., y Romano, G. 2000, "Order-theoretical ranking", *Journal of American Society for Information Science*, vol. 51, no. 7, pp. 587-601.

Cauwenberghs, G. y Poggio, T. 2000, "Incremental and decremental support vector machine learning", en *Advances in Neural Information Systems*, pp. 409-415.

Cavnar, W.B. 1994, "Using an *n*-gram-based document representation with a vector processing retrieval model", en *Proceedings of TREC-3*, pp. 269-277.

Cavnar, W.B. y Trenkle, J.M. 1994, "N-Gram-Based Text Categorization", en *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161-175.

Chakrabarti, S. 2003, *Mining the Web. Discovering Knowledge from Hypertext Data*, Morgan-Kaufmann Publishers.

Chakrabarti, S., Dom, B.E., Gibson, D., Kleinberg, J., Raghavan, P. y Rajagopalan, S. 1998a, "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text", en *Proceedings of the 7th World-Wide Web conference*, pp. 65-74.

Chakrabarti, S., Dom, B.E., Gibson, D., Kumar, R., Raghavan, P., Rajagopalan, S. y Tomkins, A. 1998b, "Experiments in topic distillation", en *Proceedings of the ACM SIGIR Workshop on Hypertext Information Retrieval on the Web*, pp. 185-193.

Chakrabarti, S., Dom, B.E., Agrawal, R. y Raghavan, P. 1998c, "Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies", *The VLDB Journal*, no. 7, pp. 163-178.

Chan, L.M. 1994, *Cataloging and Classification: An Introduction, 2nd Edition*, McGraw-Hill, Nueva York.

Chang, B. 1998, "In-place editing of web pages: Sparrow community-shared documents", *Computer Networks and ISDN Systems*, vol. 30, pp.489-498.

Chekuri, C., Goldwasser, M.H., Raghavan, P. y Upfal, E. 1997, "Web Search Using Automatic Classification", en *Sixth International World Wide Web Conference*.

Chen, X., Kwong, S. y Li, M. 1999, "A Compression Algorithm for DNA Sequences and its Applications in Genome Comparison", en *Proceedings of the 10th Workshop on Genome Informatics (GIW'99)*, pp. 51-61.

China Internet Network Information Center (CNNIC). 2003, *12th Statistical Survey on the Internet Development in China (2003/7)*.

Clark, K.G. (ed), 2004, *RDF Data Access Use Cases and Requirements*, [Online], Borrador de trabajo, Consorcio W3, Disponible en: `<http://www.w3.org/TR/rdf-dawg-uc/>` [30 Junio 2005].

Cleverdon, C., Mills, J. y Keen, E.M. 1966, *Aslib Cranfield Research Project Volume 2: Factors Determining the Performance of Indexing Systems*, ASLIB, Reino Unido.

Cleverdon, C.W. 1962, *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*, College of Aeronautics, Reino Unido (citado por Harman, D. 1993, "Overview of the First Text REtrieval Conference (TREC-1)", en *NIST Special Publication 500-207: The First Text REtrieval Conference TREC-1*).

Cohen, J.D. 1995, "Highlights: Language and Domain-Independent Automatic Indexing Terms for Abstracting", *JASIS*, 46(3), pp. 162-174.

Cohen, W. 1996, "Learning rules that classify email", en *Proceedings of the IEEE Spring Symposium on Machine Learning for Information Access*.

Cohen, W. y Hirsh, H. 1998, "Joins that generalize: Text categorization using WHIRL", en *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pp. 169-173.

Collobert, R., Bengio, S. y Bengio, Y. 2002, "Parallel mixture of SVMs for very large scale problems", *Neural Computation*, vol. 14, no. 5, pp. 1105-1114.

Conroy, J.M., Schlesinger, J.D., O'Leary, D.P. y Okurowski, M.E. 2001, "Using HMM and Logistic Regression to Generate Extract Summaries for DUC", en *Proceedings of the 1st Document Understanding Conference*.

Cowie, J., Mahesh, K., Nirenburg, S. y Zajac, R. 1998, "MINDS - Multi-lingual INteractive Document Summarization", en *AAAI 1998 Spring Symposium Series*.

Cranor, L.F. y LaMacchia, B.A. 1998, "Spam!", *CACM*, vol. 41, no. 8, pp. 74-83.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K. y Slattery, S. 1998, "Learning to Extract Symbolic Knowledge from the World Wide Web", en *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI98)*.

Crestani, F. y van Rijsbergen, C.J. 1998, "A study of probability kinematics in information retrieval", *ACM Transactions on Information Systems*, vol. 16, no. 3, pp. 225-255.

Croft, W.B. y Harper, D.J. 1979, "Using probabilistic models of document retrieval without relevance information", *Journal of Documentation*, vol. 35, pp. 285-295.

Croft, W.B. y Thompson, R.H. 1987, "I3R: A new approach to the design of document retrieval systems", *Journal of the American Society for Information Science*, vol. 38, no. 6, pp. 389-404.

Cross, V. 1994, "Fuzzy Information Retrieval", *Journal of Intelligent Information Systems*, vol. 3, pp. 29-56.

D'Amore, R., Mah, C.P. 1985, "One-time complete indexing of text: Theory and practice", en *Proceedings of SIGIR 1985*, pp. 155-164.

Damashek, M. 1995, *Gauging Similarity via N-Grams: Language-Independent Sorting, Categorization, and Retrieval of Text*. Departamento de Defensa, EE.UU.

David, C. 2003, *Information Society Statistics – PCs, Internet and mobile phone usage in the EU*, [Online], Eurostat, Disponible en: <http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-NP-03-015/EN/KS-NP-03-015-EN.PDF> [30 Junio 2005].

Davison, B.D. 2000a, "Recognizing Nepotistic Links on the Web", en *Proceedings of AAAI-2000 Workshop on Artificial Intelligence for Web Search*, pp. 23-28.

Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W. y Harshman, R.A. 1990, "Indexing by latent semantic analysis", *Journal of the Society for Information Science*, vol. 41, no. 6, pp. 391-407.

DeJong, G. 1982, "An Overview of the FRUMP System", en Lehnert, W.G. y Ringle, M.H. (Eds.) *Strategies for Natural Language Processing*, pp. 149-176.

Denker, G., Hobbs, J.R., Martin, D., Narayanan, S. y Waldinger, R. 2001, "Accessing Information and Services on the DAML-Enabled Web", en *Proceedings of the Second International Workshop Semantic Web*.

Denning, P.J. 1982, "Electronic junk", *CACM*, vol. 25, no. 3, pp. 163-165.

Domingos, P. y Pazzani, M. 1997, "On the optimality of the simple Bayesian classifier under zero one loss", *Machine Learning*, vol. 29, no. 2-3, pp. 103-130.

Doran, W., Stokes, N., Carthy, J. y Dunnion, J. 2004, "Assessing the impact of lexical chain scoring methods and sentence extraction schemes on summarization", *Proceedings of the 5th International conference on Intelligent Text Processing and Computational Linguistics CICLing-2004*.

Doyle, L.B. 1959, "Programmed interpretation of text as a basis for information retrieval systems", en *Proceedings of the Western Joint Computer Conference*, San Francisco, pp. 60-63 (citado por van Rijsbergen, C.J. 1979, *Information Retrieval, 2nd Edition*. Butterworth-Heinemann, EE.UU.)

Doyle, L.B. 1965, "Some compromises between word grouping and document grouping", en Stevens *et al.* (Eds.): *Statistical Association Methods for Mechanized Documentation*, National Bureau of Standards, EE.UU., pp. 15-24 (citado por van Rijsbergen, C.J. 1979, *Information Retrieval, 2nd Edition*. Butterworth-Heinemann, EE.UU.)

Drucker, H., Wu, D. y Vapnik, V. 1999, "Support vector machines for spam categorization", *IEEE transactions on Neural Networks*, vol. 10, no. 5, pp. 1048-1055.

Duda, R.O. y Hart, P.E. 1973, *Pattern Classification and Scene Analysis*, John Wiley and Sons, Inc. Citado en Jain, A.K., Murty, M.N. y Flynn, P.J. 1999, "Data Clustering: a review", *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323.

Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S. y Harshman, R. 1988, "Using Latent Semantic Analysis to improve access to textual information", en *Human Factors in Computing Systems, CHI'88 Conference Proceedings,* pp. 281-285.

Dumais, S.T., Platt, J., Heckerman, D. y Sahami, M. 1998, "Inductive learning algorithms and representations for text categorization", en *Proceedings of ACM-CIKM98*, pp. 148-155.

Dunlavy, D.M., Conroy, J.M., Schlesinger, J.D., Goodman, S.A., Okurowski, M.E., O'Leary, D.P. y van Halteren, H. 2003, "Performance of a three-stage system for multi-document summarization", en *Proceedings of the Document Understanding Conference (DUC)*.

Dunning, T. 1993, "Accurate methods for the statistics of surprise and coincidence", en *Computational Linguistics*, vol. 19, no. 1, pp. 61-74.

Dunning, T. 1994, *Statistical identification of language*, Informe técnico, New Mexico State University.

Dyen, I., Kruskal, J.B. y Black, P. 1992, "An Indoeuropean Classification: a lexicostatistical experiment", *Transactions of the American Philosophical Society*, vol. 82, no. 5.

EAGLES, 1996, *Preliminary recommendations on corpus typology. EAG-TCWG-CTYP/P*. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

Edmundson, H.P. 1969, "New methods in automatic abstracting", *Journal of the ACM*, vol. 16, no. 2, pp. 264-285.

Erkan, G. y Radev, D. 2004a, "LexPageRank: Prestige in multi- document text summarization", en *Proceedings of EMNLP*.

Erkan, G. y Radev, D. 2004b, "The University of Michigan at DUC 2004", en *DUC 2004*.

Evans, D.K. y Klavans, J.L. 2003, *A platform for multilingual news summarization*, [Online], Informe Técnico, University of Columbia, Disponible en : <http://www.cs.columbia.edu/~library/TR-repository/reports/reports-2003/cucs-014-03.pdf> [30 Junio 2005]

Evans, D.K., Klavans, J.L. y McKeown, K.R. 2004, "Columbia Newsblaster: Multilingual News Summarization on the Web", en *Proceedings of HLT-NAACL 2004, Demonstrations*, pp. 1-4.

Evans, S.N., Ringe, D. y Warnow, T. 2004, "Inference of divergence times as a statistical inverse problem", *Phylogenetic Methods and the Prehistory of Languages*. Cambridge, Reino Unido.

Fan, W., Gordon, D. y Pathak, P. 2004a, "A generic ranking function discovery framework by genetic programming for information retrieval", *Information Processing and Management*, vol. 40, no. 4, pp. 587-602.

Fan, W., Gordon, D. y Pathak, P. 2004b, "Discovery of context-specific ranking functions for effective information retrieval using genetic programming", *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 523-527.

Fensel, D., Angele, J., Decker, S., Erdmann, M., Schnurr, H.P., Staab, S., Studer, R. y Witt, A. 1999, "On2broker: Semantic-based access to information sources at the WWW", en *Proceedings of the World Conference on the WWW and Internet (WebNet 99)*.

Fensel, D., Decker, S., Erdmann, M. y Studer, R. 1998, "Ontobroker: Or How to Enable Intelligent Access to the WWW", en *Proceedings of the 11th Workshop on Knowledge Acquisition, Modeling, and Management*.

Ferreira da Silva, J., Pereira Lopes, G. 1999, "A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora", en *Proceedings of MOL6*, pp. 369-381.

Filo, D. y Yang, J. 1994, *Yahoo!*, [Online], Yahoo! Inc., Disponible en: <http://www.yahoo.com> [30 Junio 2005].

Flake, G.W., Pennock, D.M. y Fain, D.C. 2003, "The Self-Organized Web: The Yin to the Semantic Web's Yang", *IEEE Intelligent Systems*, vol. 18, no. 4, pp. 75-77.

Fletcher, J. 1994, *Internet Robots – Structure from Anarchy?*, [Online], Internet Archive, Disponible en: <http://web.archive.org/web/19990116235019/http://lorne.stir.ac.uk/~jf1/papers/signidr.html> [30 Junio 2005].

Florescu, D., Levy, A. y Mendelzon, A. 1998, "Database Techniques for the World-Wide Web: A Survey", *ACM SIGMOD Record*, vol. 27, no. 3, pp. 59-74.

Foltz, P.W. 1990, "Using Latent Semantic Indexing for Information Filtering", en *Proceedings of the ACM Conference on Office Information Systems*, pp. 40-47.

Foltz, P.W. y Dumais, S.T. 1992, "Personalized Information Delivery: An Analysis of Information Filtering Methods", *Communications of the ACM*, vol. 35, no. 12, pp. 51-60.

Fox, E. 1983. *Characteristics of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts*, Informe técnico, Cornell University.

Fuentes, M., Massot, M., Rodríguez, H. y Alonso, L. 2003, "Mixed approach to headline extraction for DUC 2003", en *Proceedings of DUC 2003*.

Fukumoto, F., Suzuki, Y. y Fukumoto, J. 1997, "An Automatic Extraction of Key Paragraphs Based on Context Dependency", en *Proceedings of the 5th International on Applied Natural Language Processing*.

Fum, D., Guida, G. y Tasso, C. 1985, "Evaluating Importance: A Step Towards Text Summarization", en *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pp. 840-844.

Furnas, G.W., Landauer, T.K., Gómez, L.M. y Dumais, S.T. 1987, "The vocabulary problem in human-system communication", *Communications of the ACM*, vol. 30, no. 11, pp. 964-971.

Fussler, H.H. 1949, "Characteristics of the research literature used by chemists and physicists in the United States" en *The origins of information science*, 1987, ed. A.J. Meadows, Taylor Graham Publishing.

Galliers, J.R. y Spärck-Jones, K. 1993, *Evaluating Natural Language Processing Systems*, Informe Técnico, Computer Laboratory, University of Cambridge.

Garfield, E. 1972, "Citation Analysis as a Tool in Journal Evaluation", *Science*, vol. 178, pp. 471-479.

Gawronska, B. 2002, "Employing Cognitive Notions in Multilingual Summarization of News Reports", en *Proceedings of NLULP-02 (The 7th International Workshop on Natural Language Understanding and Logic Programming)*.

Gayo Avello, D. y Álvarez Gutiérrez, D. 2002, "The Cooperative Web: A Complement to the Semantic Web", en *Proceedings of the 26th Annual International Computer Software & Applications Conference (COMPSAC'02)*, pp. 179-183.

Gayo Avello, D., Álvarez Gutiérrez, D. y Gayo Avello, J. 2004a, "Naïve Algorithms for Keyphrase Extraction and Text Summarization from a Single Document Inspired by the Protein Biosyntehis Process", en A.J. Ijspeert *et al.* (Eds.): *BioADIT 2004*, LNCS 3141, pp. 440-455.

Gayo Avello, D., Álvarez Gutiérrez, D. y Gayo Avello, J. 2004b, "One Size Fits All? A Simple Technique to Perform Several NLP Tasks", en J.L. Vicedo *et al.* (Eds.): *EsTAL 2004*, LNAI 3230, pp. 267-278.

Gayo Avello, D., Álvarez Gutiérrez, D. y Gayo Avello, J. 2004c, "Application of variable length *n*-grams to monolingual and bilingual information retrieval", en Peters, C. y Borri, F. (Eds.): *Working Notes for the CLEF 2004 Workshop*.

Goddard, I. y Campbell, L. 1994, "The History and Classification of American Indian Languages: What are the Implications for the Peopling of the Americas?", en Bonnichsen, R. y Steele, D.G. (Eds.): *Method and Theory for Investigating the Peopling of the Americas*.

Goldberg, D., Nichols, D., Oki, B.M. y Terry, D. 1992, "Using Collaborative Filtering to Weave an Information Tapestry", *Communications of the ACM*, vol. 35, no. 12, pp. 61-70.

Goldman, C.V., Langer, A. y Rosenschein, J.S. 1996, "Musag: an agent that learns what you mean", en *Proceedings of The First International Conference on The Practical Application of Intelligent Agents and Multi Agent Technology (PAAM '96)*, pp. 311-329.

Graham, L. y Metaxas, P.T. 2003, "Of Course it's True; I Saw It on the Internet! Critical Thinking in the Internet Era", *Communications of the ACM*, vol. 46, no. 5, pp. 71-75.

Gray, M. 1995, *Measuring the Growth of the Web – June 1993 to June 1995*, [Online], MIT, Disponible en: <http://www.mit.edu/people/mkgray/growth> [30 Junio 2005].

Gray, R.D. y Atkinson, Q.D. 2003, Language-tree divergence times support the Anatolian theory of Indo-European origin, *Nature*, 426, pp. 435-439.

Greenberg, J.H. 1966, *The Languages of Africa*, Bloomington, Indiana University.

Greenberg, S. y Roseman, M. 1996, "GroupWeb: A WWW Browser as Real Time Groupware", en *Proceedings of ACM SIGCHI'96 Conference on Human Factors in Computing System*.

Grefenstette, G. 1995, "Comparing two language identification schemes", en *Proceedings of 3rd International Conference on Statistical Analysis of Textual Data*.

Griffith, B.C. 1980, *Key Papers in Information Science*, American Society for Information Science, EE.UU.

Gross, P.L.K. y Gross, E.M. 1927, "College libraries and chemical education", en *The origins of information science*, 1987, ed. A.J. Meadows, Taylor Graham Publishing.

Gruber, T.R. 1993, "A Translation Approach to Portable Ontology Specifications", *Knowledge Acquisition*, vol. 5, no. 2, pp. 199-220.

Guha, R.V., Lassila, O., Miller, E. y Brickley, D. 1998, "Enabling inferencing", [Online], en *W3C Query Languages Workshop (QL'98)*, Disponible en: <http://www.w3.org/TandS/QL/QL98/pp/enabling.html> [30 Junio 2005].

Han, E., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., Mobasher, B. y Moore, J. 1998, "WebACE: A Web Agent for Document Categorization and Exploration", en *Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98)*, pp. 408-415.

Han, E., Karypis, G. y Kumar, V. 1999, *Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification*, Informe técnico, Universidad de Minnesota.

Hardy, H., Shimizu, N., Strzalkowski, T., Ting, L., Zhang, X. y Wise, B. 2001, "Cross-Document Summarization by Concept Classification", *Document Understanding Conference, a SIGIR'01 workshop*.

Harman, D. 1993, "Overview of the First Text REtrieval Conference (TREC-1)", en *NIST Special Publication 500-207: The First Text REtrieval Conference TREC-1*.

Harman, D. 1996, "Overview of the Fourth Text REtrieval Conference (TREC-4)", en *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*.

Harper, D. y van Rijsbergen, C.J. 1978, "An evaluation of feedback in document retrieval using co-occurrence data", *Journal of Documentation*, vol. 34, pp. 189-216.

Hearst, M.A. 1994, "Multi-Paragraph Segmentation of Expository Text", en *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, pp. 9-16.

Hearst, M.A. y Karadi, C. 1997, "Cat-a-Cone:An Interactive Interface for Specifying Searching and Viewing Retrieval Results using a Large Category Hierarchy", en *Proceedings of the 20th Annual International ACM/SIGIR Conference*.

Henderson, M.M. 1999, "Examples of Early Nonconventional Technical Information Systems", en *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*, eds. Bowden, M.E., Hahn, T.B. y Willians R.V., Information Today Inc., pp. 169-176.

Hersh, W., Buckley, C., Leone, T. y Hickam, D. 1994, "OHSUMED: An interactive retrieval evaluation and new large test collection for research", en *Proc. of SIGIR'94*, pp. 192-201.

Hirao, T., Sasaki, Y., Isozaki, H. y Maeda, E. 2002, "NTT's Text Summarization System for DUC 2002", en *Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 40th Meeting of the Association for Computational Linguistics*.

Hirao, T., Suzuki, J., Isozaki, H. y Maeda, E. 2003, "NTT's Multiple Summarization System for DUC2003", en *Proc. of the 3th Document Understanding Conference (DUC)*.

Honkela, T., Kaski, S., Lagus, K., y Kohonen, T. 1996, *Newsgroup exploration with WEBSOM method and browsing interface*, Informe técnico, Universidad Tecnológica de Helsinki.

Horrocks, I., Fensel, D., Broekstra, J., Decker, S., Erdmann, M., Goble, C., van Harmelen, F., Klein, M., Staab, S., Studer, R. y Motta, E. 2000, *The Ontology Inference Layer OIL. Technical report*, [Online], Informe técnico, On-To-Knowledge, Disponible en: <http://www.ontoknowledge.org/oil/TR/oil.long.html> [30 Junio 2005].

Hovy, E. (Ed.) 1999, "Cross-lingual Information Extraction and Automated Text Summarization", en *Multilingual Information Management: Current Levels and Future Abilities* (Hovy, E. *et al*, eds.), [Online], US National Science Fundation, Disponible en: <http://www-2.cs.cmu.edu/~ref/mlim/> [30 Junio 2005]

Hovy, E. y Lin, C.Y. 1997, "Automated Text Summarization in SUMMARIST", en *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, pp. 18-24.

Hovy, E. y Lin, C.Y. 1998, "Automated Text Summarization and the SUMMARIST System", en *Proceedings of the TIPSTER Text Program Phase III final report*, pp. 197-214.

Huffman, S.M. 1998, *The Genetic Classification of Languages by n-Gram Analysis: A Computational Technique*, tesis doctoral, Universidad de Georgetown.

Hull, D.A. 1994, "Improving text retrieval for the routing problem using latent semantic indexing", en Croft, W.B. y van Rijsbergen, C.J. (Eds.): *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pp. 282-289.

HyperText Mark-up Language, 1992, *HyperText Mark-up Language*, [Online], W3 Consortium, Disponible en: <http://www.w3.org/History/19921103-hypertext/hypertext/WWW/MarkUp/MarkUp.html> [30 Junio 2005].

IBM, 1994-2004, *IBM 1301 disk storage unit*, [Online], IBM Corporation, Disponible en: <http://www.ibm.com/ibm/history/exhibits/storage/storage_1301.html> [30 Junio 2005].

IBM, 2002, *Fifty Years Of Storage Innovation: Magnetic Tape & Beyond*, [Online], IBM Corporation, Disponible en: <http://www.ibm.com/ibm/history/exhibits/storage/storage_fifty.html> [30 Junio 2005].

Ingwersen, P. 1992, *Information Retrieval Interaction*, Taylor Graham, Reino Unido, Disponible en: <http://www.db.dk/pi/iri/> [30 Junio 2005]

IPA (*International Phonetic Association*) 1999, *Handbook of the International Phonetic Association : A Guide to the Use of the International Phonetic Alphabet*, Cambridge University Press, Reino Unido.

Ittner, D.J., Lewis, D.D. y Ahn, D.D. 1995, "Text categorization of low quality images", en *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, pp. 301-315.

Iwayama, M. y Tokunaga, T. 1995, "Hierarchical Bayesian Clustering for Automatic Text Classification", en *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95*, pp. 1322-1327.

Jain, A.K. y Dubes, R.C. 1988, *Algorithms for Clustering Data*, Prentice Hall. Citado en Zhao, Y. y Karypis, G. 2002, *Criterion Functions for Document Clustering: Experiments and Analysis*, Informe técnico, Universidad de Minnesota.

Jain, A.K., Murty, M.N. y Flynn, P.J. 1999, "Data Clustering: a review", *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323.

Jalam, R. y Teytaud, O. 2001, "Kernel-based text categorization", en *International Joint Conference on Neural Networks (IJCNN'2001)*.

Jansen, B.J. y Spink, A. 2003, "An Analysis of Web Documents Retrieved and Viewed", *The 4th International Conference on Internet Computing*, pp. 65-69.

Jansen, B.J., Spink, A., Bateman, J. y Saracevic, T. 1998, "Real life information retrieval: A study of user queries on the web", *SIGIR Forum*, vol. 32, no. 1, pp. 5-17.

Jaoua, M. y Ben Hamadou, A. 2003, "Automatic Text Summarization of Scientific Articles Based on Classification of Extract's Population", *CICLing 2003* (citado por Alfonseca, E. *et al.* 2004, "Description of the UAM system for generating very short summaries at DUC-2004", *NAACL Text Summarization Workshop and Document Understanding Conference 2004*).

Jardine, N. y van Rijsbergen, C.J. 1971, "The use of hierarchic clustering in information retrieval", *Information Storage and Retrieval*, vol. 7, pp. 217-240.

Järvelin, K. y Vakkari, P. 1992, "Content Analysis of research articles in LIS 1965-1985", en Cronin, B. y Vakkari, P. (Eds.): *Conceptions of Library and Information Science. Proceedings of First CoLIS Conference* (citado por *Ingwersen*, P. 1992, Information Retrieval Interaction, Taylor Graham, Reino Unido).

Jarvis, R.A. y Patrick, E.A. 1973, "Clustering Using a Similarity Measure Based on Shared Near Neighbors", *IEEE Transactions on Computers*, pp. 1025-1034.

Jing, H. 2000, "Sentence reduction for automatic text summarization", en *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*.

Jing, H. y McKeown, K.R. 2000, "Cut and Paste-Based Text Summarization", en *Proceedings of the 6th Applied Natural Language Processing Conference and the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 178-185.

Jing, H., Barzilay, R., McKeown, K. y Elhadad, M. 1998, "Summarization evaluation methods: Experiments and analysis", en *Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pp. 60-68.

Joachims, T. 1997, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization", en *Proceedings of the 14th International Conference on Machine Learning*, pp.143-151.

Joachims, T. 1998, "Text categorization with support vector machines: Learning with many relevant features", en *Proceedings of the European Conference on Machine Learning*, pp. 137-142.

Johnson, T. 1993, *Re: WWW Information Discovery Tools*, [Online], University of Calgary, Disponible en: <http://ksi.cpsc.ucalgary.ca/archives/WWW-TALK/www-talk-1993q2.messages/30.html> [30 Junio 2005].

Kando, N. 1999, "Preface", en *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, [Online], Disponible en: <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings> [30 Junio 2005]

Kantor, P.B., Boros, E., Melamed, B., Menkov, V., Shapira, B. y Neu, D.J. 2000, "Capturing human intelligence in the Net", *Communications of the ACM*, vol. 43, no. 8, pp. 112-115.

Kantor, P.B., Voorhees, E.M. 2000, "The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text", *Information Retrieval*, vol. 2, no. 3, pp. 165-176.

Karvounarakis, G., Christophides, V., Plexousakis, D. y Alexaki, S. 2001, "Querying RDF Descriptions for Community Web Portals", *The French National Conference on Databases*.

Karypis, G. y Han, E. 2000, *Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization*, Informe técnico, University of Minnesota.

Kaufman, L. 1998, "Solving the Quadratic Programming Problem Arising in Support Vector Classification", en *Advances in Kernel Methods - Support Vector Learning*, pp. 147-167. Citado por Platt, J.C. 1999, "Using Analytic QP and Sparseness to Speed Training of Support Vector Machines", *NIPS 11*, pp. 557-563.

Kearns, M. 1988, *Thoughts on Hypothesis Boosting*, manuscrito no publicado, [Online], Disponible en: <http://www.cis.upenn.edu/~mkearns/papers/boostnote.pdf> [30 Junio 2005].

Kessler, B. 1995, "Computational Dialectology in Irish Gaelic", en *Proceedings of the European Association for Computational Linguistics*, pp. 60-67.

King, B. 1963, "Step-wise clustering procedures", *Journal of the American Statistical Association*, 69, pp. 86-101.

Kleinberg, J.M. 1998, "Authoritative sources in a hyperlinked environment", en *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pp. 668-677.

Knight, K. y Marcu, D. 2000, "Statistics-based summarization – step one: Sentence compression", en *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, pp. 703-710.

Koehn, P., *Europarl: A Multilingual Corpus for Evaluation of Machine Translation*, manuscrito no publicado, [Online], Disponible en: <http://people.csail.mit.edu/~koehn/publications/europarl.ps> [30 Junio 2005]

Kohonen, T. 1982, "Self-organized formation of topologically correct feature maps", *Biological Cybernetics*, 43, pp. 59-69.

Kolda, T.G. y O'Leary, D.P. 1998, "A Semidiscrete Matrix Decomposition for Latent Semantic Indexing in Information Retrieval", *ACM Trans. Information Systems*, vol. 16, pp. 322-346.

Koller, D. y Sahami, M. 1997, "Hierarchically classifying documents using very few words", en *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pp. 170-178.

Koller, D. y Sahami, M. 1997, "Hierarchically classifying documents using very few words", en *Proc. of the 14th Int. Conf. on Machine Learning*, pp. 170-178.

Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R. y Riedl, J. 1997, "GroupLens: Applying Collaborative Filtering to Usenet News", *Communications of the ACM*, vol. 40, no. 3, pp. 77-87.

Koster, M. 1994, "ALIWEB – Archie-Like Indexing in the WEB", *Computer Networks and ISDN Systems*, vol. 27, no. 2, pp. 175-182.

Kovács, L. y Micsik, A. 2000, "The Collaborative Web", *ERCIM News*, no. 41.

Kraaij, W., Spitters, M. y Hulth, A. 2002, "Headline extraction based on a combination of uni- and multidocument summarization techniques", en *Proceedings of the ACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2002)*.

Kraaij, W., Spitters, M. y van der Heijden, M. 2001, "Combining a mixture language model and Naive Bayes for multi-document summarisation", en *Proceedings of the DUC 2001 workshop (SIGIR 2001)*.

Kraft, D.H. y Buell, D.A. 1983, "Fuzzy sets and generalized Boolean retrieval systems", *International Journal of Man-Machine Studies*, vol. 19, pp. 45-56.

Kruijff, G.M., 2002, *Learning linearization rules from treebanks*, [Online], Disponible en: <http://www.cs.haifa.ac.il/~shuly/fg02/kruijff.pdf> [30 Junio 2005].

Kuhner, M.K. y Felsenstein, J. 1994, "A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates", *Molecular Biology and Evolution*, vol. 11, no. 3, pp. 459-468.

Kupiec J., Pedersen J. y Chen F. 1995, "A Trainable Document Summarizer", en *Proceedings of the ACM SIGIR conference*, pp. 68-73.

Kwok, J.T.Y. 1998, "Automated text categorization using support vector machine", en *Proceedings of the International Conference on Neural Information Processing (ICONIP)*, pp. 347-351.

Lal, P. y Ruger, S. 2002, "Extract-based Summarization with Simplification", en *Proceedings of the ACL Workshop on Text Summarisation / DUC 2002*.

Lam, S. y Lee, D. 1999, "Feature reduction for neural network based text categorization", en *6th Conference On Database Systems For Advanced Applications*, pp. 195-202.

Lancaster, F.W. 1968, *Information Retrieval Systems: Characteristics, Testing and Evaluation*, Wiley, EE.UU. (citado por van Rijsbergen, C.J. 1979, *Information Retrieval, 2nd Edition*, Butterworth-Heinemann, EE.UU.)

Langer, S. 2001, "Natural languages on the Word Wide Web", [Online], en *Bulag. Revue annuelle*, pp. 89-100, disponible en: <http://www.cis.uni-muenchen.de/people/langer/veroeffentlichungen/bulag.pdf> [30 Junio 2005]

Larkey, L.S. y Croft, W.B. 1996, "Combining classifiers in text categorization", en *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 289-297.

Larocca Neto, J., Santos, A.D., Kaestner, C.A.A. y Freitas, A.A. 2000, "Document Clustering and Text Summarization", en *Proceedings of the 4th International Conference Practical Applications of Knowledge Discovery and Data Mining*, pp. 41-55.

Lawrence, S. y Giles, C.E. 1998, "Searching the World Wide Web", *Science*, vol. 280, no. 3, pp. 98-100.

Lees, R.B. 1953, "On the basis of glottochronology", *Language*, 29, pp. 113-127.

Lenci, A., Bartolini, R., Calzolari, N., Agua, A., Busemann, S., Cartier, E., Chevreau, K. y Coch, J. "Multilingual summarization by integrating linguistic resources in the mlis-musi project", en *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02).*

Lenhart, A., Fallows, D. y Horrigan, J. 2004, *Content Creation Online*, Informe, Pew Internet & American Life Project.

Lerman, I.C. 1970, *Les Bases de la Classification Automatique*, Gauthier-Villars, Paris. Citado en van Rijsbergen, C.J. 1979, *Information Retrieval, 2nd Edition*. Butterworth-Heinemann, EE.UU.

Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals, (traducción inglesa del ruso), *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707-710.

Lewis, D. y Catlett, J. 1994. "Heterogeneous uncertainty sampling for supervised learning", en *Proceedings of the 11th International Conference on Machine Learning*, pp. 148-156.

Lewis, D.D. 1991, "Evaluating text categorization", en *Proceedings of Speech and Natural Language Workshop*, pp. 312-318.

Lewis, P.A.W, Baxendale, P.B. y Bennett, J.L. 1967, "Statistical Discrimination of the Synonymy/Antonymy Relationship Between Words", *Journal of the ACM*, vol. 14, no. 1, pp. 20-44.

Li, H. y Yamanishi, K. 1999, "Text classification using esc-based stochastic decision lists", en *8th ACM International Conference on Information and Knowledge Management (CIKM99)*, pp. 122-130.

Li, M., Chen, X., Li, X., Ma, B. y Vitányi, P. 2004, "The Similarity Metric", *IEEE Transactions on Information Theory*, vol. 50, pp. 3250-3264.

Li, W.S., Vu, Q., Chang, E., Agrawal, D., Hara, Y. y Takano, H. 1999, "PowerBookmarks: A System for Personalizable Web Information Organization, Sharing, and Management", en *Proceedings of the Eighth International World-Wide Web Conference.*

Lieberman, H. 1995, "Letizia: An Agent That Assists Web Browsing", en *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 924-929.

Lin, C.Y. 1999, "Machine Translation for Information Access across the Language Barrier: the MuST System", en *Proceedings of the Machine Translation Summit VII.*

Lin, C.Y. 2003, "Improving Summarization Performance by Sentence Compression – A Pilot Study", en *Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages.*

Lin, C.Y. 2004a, "Rouge: A Package for Automatic Evaluation of Summaries", en *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74-81.

Lin, C.Y. 2004b, "Looking for a Few Good Metrics: Automatic Summarization Evaluation – How Many Samples Are Enough?", en *Proceedings of NTCIR Workshop 4.*

Lin, C.Y. y Hovy, E.H. 1997, "Identifying Topics by Position", en *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP).*

Lin, C.Y. y Hovy, E.H. 2000, "The automated acquisition of topic signatures for text summarization", en *Proceedings of the 18th COLING Conference.*

Lin, C.Y. y Hovy, E.H. 2001, "NeATS: A multidocument summarizer", en *Proceedings of the Document Understanding Conference (DUC01)*.

Lin, C.Y. y Hovy, E.H. 2002a, "NeATS in DUC 2002", en *Proceedings of the Document Understanding Conference (DUC02)*.

Lin, C.Y. y Hovy, E.H. 2002b, "Manual and Automatic Evaluation of Summaries", en *Proceedings of the Workshop on Automatic Summarization (including DUC 2002)*, pp. 45-51.

Lin, C.Y. y Hovy, E.H. 2003, "Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics", en *Proceedings of the Human Technology Conference 2003 (HLT-NAACL-2003)*.

Lin, C.Y. y Hovy, E.H. 2003, "Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics", en *Proceedings of the Human Technology Conference 2003 (HLT-NAACL-2003)*.

Lotka, A.J. 1926, "The frequency distribution of scientific productivity", en *The origins of information science*, 1987, ed. A.J. Meadows, Taylor Graham Publishing.

Luhn, H.P. 1957, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information", *IBM Journal of Research and Development*, vol. 1, no. 4, pp. 309-317.

Luhn, H.P. 1958, "The Automatic Creation of Literature Abstracts", *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159-165.

Luhn, H.P. 1959, *Keyword-in-context index for technical literature (KWIC)*, IBM Advanced System Development Division, RC-127 (citado por Ohlman, H. 1999, "Mechanical Indexing: A Personal Remembrance", en Bowden, M.E., Bellardo Hahn, T. y Williams, R.V. (Eds.): *Proc. 1998 Conf. History and Heritage of Science Information Systems*, ASIS Monograph Series, Information Today, EE.UU., pp. 187-192).

Luke, S., Spector, L. y Rager, D. 1996, "Ontology-Based Knowledge Discovery on the World-Wide Web", en *Working Notes of the Workshop on Internet-Based Information Systems at the 13th National Conference on Artificial Intelligence (AAAI96)*.

Maarek, Y.S. y Ben Shaul, I.Z. 1996, "Automatically Organizing Bookmarks per Contents", en *Proceedings of the 5th International World Wide Web Conference*.

Maes, P. 1994, "Agents that Reduce Work and Information Overload", *Communications of the ACM*, vol. 37, no. 7, pp. 811-821.

Mani, I. y Maybury, M.T. (Eds.) 1999, *Advances in Automatic Text Summarization*, The MIT Press.

Mani, I., Gates, B. y Bloedorn, E. 1999, "Improving Summaries by Revising Them", en *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 558-565.

Mani, I., House, D., Klein, G., Hirschman, L., Obsrt, L., Firmin, T., Chrzanowski, M. y Sundheim, B. 1998, *The TIPSTER SUMMAC Text Summarization Evaluation: Final Report*, Informe técnico, [Online], Disponible en: <http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/summac-final-report-part2.ps> [30 Junio 2005]

Marchiori, M. y Saarela, J. 1998, "Query + Metadata + Logic = Metalog", en *Proceedings of Query Languages Workshop*.

Marchiori, M. y Saarela, J. 1999a, *Let's Reason on the Web: Metalog*, [Online], Consorcio W3C, Disponible en: <http://www.w3.org/RDF/Metalog/metalog990224.html> [30 Junio 2005].

Marchiori, M. y Saarela, J. 1999a, *Towards the Semantic Web: Metalog*, [Online], Consorcio W3C, Disponible en: <http://www.w3.org/RDF/Metalog/CIKM-050299.html> [30 Junio 2005].

Maron, M.E. y Kuhns, K.L. 1960, "On relevance, probabilistic indexing and information retrieval", *Journal of the ACM*, vol. 7, no. 3, pp. 216-244.

Mauldin, M.L. y Leavitt, J.R.R. 1994, "Web Agent Related Research at the Center for Machine Translation", [Online], en *Proceedings of the ACM Special Interest Group on Networked Information Discovery and Retrieval*, Disponible en: <http://web.archive.org/web/19970607125802/http://fuzine.mt.cs.cmu.edu/mlm/signidr94.html> [30 Junio 2005].

McBryan, O.A. 1994, "GENVL and WWWW: Tools for taming the Web", *Computer Networks and ISDN Systems*, vol. 27, no. 2, p. 308.

McCallum, A.K., Rosenfeld, R., Mitchell, T.M. y Ng, A.Y. 1998, "Improving text classification by shrinkage in a hierarchy of classes", en *Proceedings of ICML-98, 15th International Conference on Machine Learning*, pp. 359-367.

McCulloch, W.S. y Pitts, W. 1943, "A logical calculus of ideas immanent in nervous activity", *Bulletin of Mathematical Biophysics*, 5, pp. 115-133.

McGuinness, D.L., Fikes, R., Stein, L.A. y Hendler, J. 2000, "DAML-ONT: An Ontology Language for the Semantic Web", en *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*, Fensel, D., Hendler, J., Lieberman, H. y Wahlster, W. (eds), pp. 65-94.

McKeown, K.R., Barzilay, R., Blair-Goldensohn, S., Evans, D., Hatzivassiloglou, V., Klavans, J., Nenkova, A., Schiffman, B. y Sigelman, S. 2002, "The Columbia Multi-Document Summarizer", en *Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 40th Meeting of the Association for Computational Linguistics*.

McKeown, K.R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Teufel, S., Kan, Y.M. y Schiffman, B. 2001, "Columbia Multi-Document Summarization: Approach and Evaluation", en *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

McNamee, P. y Mayfield, J. 2003, "JHU/APL Experiments in Tokenization and Non-Word Translation", *Working Notes for the CLEF 2003 Workshop*. 21-22 August, Trondheim, Norway.

McNamee, P. y Mayfield, J. 2004, "Cross-Language Retrieval Using HAIRCUT for CLEF 2004", en *Working Notes for the CLEF 2004 Workshop (Volume I: Papers)*, pp. 31-37.

Menczer, F. y Belew, R.K. 1998, "Adaptive Information Agents in Distributed Textual Environments", en *Proceedings of the 2nd International Conference on Autonomous Agents*, pp. 157-164.

Menczer, F., Belew, R.K. y Willuhn, W. 1995, "Artificial Life Applied to Adaptive Information Agents", en *Proceedings of AAAI Spring Symposium on Information Gathering*.

Merkl, D. y Rauber, A. 1998, "CIA's View of the World and What Neural Networks Learn from It: A Comparison of Geographical Document Space Representation Metaphors", en Quirchmayr, G. *et al.* (Eds.): *Database and Expert Systems Applications, 9th International Conference*, LNCS 1460, pp. 816-825

Miller, L., Seaborne, A. y Reggiori, A. 2002, "Three Implementations of SquishQL, a Simple RDF Query Language", en *Proceedings of 1st International Semantic Web Conference. ISWC2002*.

Miller, M. 1998, *The GnuHoo BooBoo*, [Online], Slashdot, Disponible en: `<http://slashdot.org/article.pl?sid=98/06/23/0849239&mode=nested&tid=95>` [30 Junio 2005].

Ministry of Public Management, Home Affairs, Posts and Telecommunications (MPHPT). 2001, *Number of Internet Users in CY2000 Reaches 47,080,000*.

Minsky, M. y Papert, S. 1969, *Perceptrons: an introduction to computational geometry*. MIT Press. Citado en Elkan, C. 1997, *Boosting and Naive Bayesian learning*, Informe Técnico, University of California, San Diego.

Mitra, M., Singhal, A. y Buckley, C. 1997, "Automatic text summarization by paragraph extraction", en *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pp. 31-36.

Mitra, M., Singhal, A. y Buckley, C. 1998, "Improving automatic query expansion", en *ACM SIGIR'98*, pp. 20-27.

Morita, M. y Shinoda, Y. 1994, "Information filtering based on user behavior analysis and best match text retrieval", en *Proceedings of the 17th Annual International Retrieval*, pp. 272-281.

Mosteller, F. y Wallace, D.L. 1964, "Inference and Disputed Authorship: The Federalist Series", en *Behavioral Science: Quantitative Methods*, Citado en Fung, G. 2003, "The Disputed Federalist Papers: SVM and Feature Selection via Concave Minization", [Online], en *Proceedings of the 2003 Conference of Diversity in Computing*, pp. 42-46, disponible en: `<http://www.cs.wisc.edu/~gfung/federalist.pdf>` [30 Junio 2005]

Moulinier, I. y Ganascia, J.G. 1996, "Applying an existing machine learning algorithm to text categorization", en *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pp. 343-354.

Muthukrishnan, S. y Sahinalp, S.C. 2000, "Approximate nearest neighbors and sequence comparison with block operations", en *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pp. 416-424.

Nenkova, A. y Passonneau, R. 2004, "Evaluating Content Selection in Summarization: The Pyramid Method", en *Proceedings of HLT/NAACL 2004*.

Nenkova, A., Schiffman, B., Schlaiker, A., Blair-Goldensohn, S., Barzilay, R., Sigelman, S., Hatzivassiloglou, V. y McKeown, K.R. 2003, "Columbia at the DUC 2003", en *DUC03*.

Nerbonne, J. y Heeringa, W. 1997, "Measuring Dialect Distance Phonetically", en *Proceedings of the third meeting of the ACL special interest group in computational phonology*, pp. 11-18.

Ng, H., Goh, W. y Low, K. 1997, "Feature selection, perceptron learning, and a usability case study for text categorization", en *Proc. 20th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'97)*, pp. 67-73.

NIST, 2004. *DUC 2004: Documents, Tasks, and Measures*, [Online], Disponible en: <http://duc.nist.gov/duc2004/tasks.html> [30 Junio 2005].

Ohlman, H. 1957, *Permutation Indexing: Multiple-Entry Listing by Electronic Accounting Machines*, Rand Corp. System Development Div., EE.UU.

Ohlman, H. 1999, "Mechanical Indexing: A Personal Remembrance", en Bowden, M.E., Bellardo Hahn, T. y Williams, R.V. (Eds.): *Proc. 1998 Conf. History and Heritage of Science Information Systems*, ASIS Monograph Series, Information Today, EE.UU., pp. 187-192.

Ontrup, J. y Ritter, H. 2001, "Text categorization and semantic browsing with self-organizing maps on non-euclidean spaces", en *Proceedings of PKDD-01*, pp. 338-349.

Osuna, E., Freund, R. y Girosi, F. 1997, *Support vector machines: Training and applications*, informe técnico, MIT.

Page, L., Brin, S., Motwani, R. y Winograd, T. 1998, *The PageRank Citation Ranking: Bringing Order to the Web*, [Online], Stanford University, Disponible en: <http://dbpubs.stanford.edu/pub/1999-66> [30 Junio 2005].

Paice, C.D. y Jones, P.A. 1993, "The identification of important concepts in highly structured technical papers", en Korfhage, R., Rasmussen, E. y Willet, P. (Eds.) *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 69-75.

Pantel, P. y Lin, D. 1998, "SpamCop: A Spam Classification & Organization Program", en *Learning for Text Categorization: Papers from the 1998 Workshop*.

Papineni, K., Roukos, S., Ward, T. y Zhu, W., 2002, "Bleu: a Method for Automatic Evaluation of Machine Translation", en *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311-318.

Peinado, V., Artiles, J., López-Ostenero, F., Gonzalo, J. y Verdejo, F. 2004, "UNED at Image CLEF 2004: Detecting Named Entities and Noun Phrases for Automatic Query Expansion and Structuring", en Peters, C. y Borri, F. (Eds.): *Working Notes for the CLEF 2004 Workshop*.

Perry, W.M. 1993, *Re: WWW Information Discovery Tools*, [Online], University of Calgary, Disponible en: <http://ksi.cpsc.ucalgary.ca/archives/WWW-TALK/www-talk-1993q2.messages/27.html> [30 Junio 2005].

Peters, C. 2001, "Introduction", en Peters, C. (Ed.): *CLEF 2000*, LNCS 2069, pp. 1-6.

Peters, C., Clough, P.D., Jones, G.J.F., Gonzalo, J., Kluck, M. y Magnini, B. (Eds.) 2005, *Multilingual Information Access for Text, Speech and Images: Results of the 5th CLEF Evaluation Campaign*, LNCS 3491.

Pinkerton, B. 1994, "Finding what people want: Experiences with the WebCrawler", [Online], Internet Archive, en *Electronic Proceedings of the "Second World Wide Web Conference '94: Mosaic and the Web"*, NCSA, Disponible en: <http://archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/pinkerton/WebCrawler.html> [30 Junio 2005].

Pirkola, A, Keskustalo, Heikki, Leppänen, Erkka, Känsälä, Antti-Pekka and Järvelin, Kalervo, 2002, "Targeted *s*-gram matching: a novel *n*-gram matching technique for cross- and monolingual word form variants", *Information Research*, vol. 7, no. 2.

Platt, J.C. 1998, "How to implement SVMs", *IEEE Intelligent Systems*, vol. 13, no. 4.

Platt, J.C. 1999, "Using Analytic QP and Sparseness to Speed Training of Support Vector Machines", *NIPS 11*, pp. 557-563.

Porter, M.F. 1980, "An algorithm for suffix stripping", *Program*, vol. 14, no. 3, pp. 130-137.

Poser, W.J. y Campbell, L. 1992, "Indo-European Practice and Historical Methodology", en *Proceedings of the Eighteenth Annual Meeting of the Berkeley Linguistics Society*, pp. 214-236.

Postel, J. 1975, *On the Junk Mail Problem*, RFC-0706, [Online], disponible en: <http://www.faqs.org/rfcs/rfc706.html> [30 Junio 2005]

Poutsma, A. 2002, "Applying Monte Carlo Techniques to Language Identification", en *CLIN 2001 (Twelfth Meeting of Computational Linguistics in the Netherlands)*.

Prager, J.M. 1999, "Linguini: Recognition of Language in Digital Documents", en *Proceedings of the 32nd Hawaii International Conference on System Sciences*.

Prud'hommeaux, E. y Seaborne, A. 2004, *BRQL - A Query Language for RDF*, [Online], Borrador de trabajo, Consorcio W3, Disponible en: <http://www.w3.org/2001/sw/DataAccess/rq23/> [30 Junio 2005].

Putz, S. 1993, *Re: WWW Information Discovery Tools*, [Online], University of Calgary, Disponible en: <http://ksi.cpsc.ucalgary.ca/archives/WWW-TALK/www-talk-1993q2.messages/31.html> [30 Junio 2005].

Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Çelebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A. y Zhu, Z. "MEAD – A platform for multidocument multilingual text summarization", en *Proceedings of LREC 2004*.

Radev, D., Blair-Goldensohn, S. y Zhang, Z. 2001, "Experiments in single and multi-document summarization using MEAD", en *First Document Understanding Conference*.

Radev, D., Otterbacher, J., Qi, H. y Tam, D. 2003, "MEAD ReDUCs: Michigan at DUC 2003", en *DUC03*.

Radev, D., Teufe, S., Saggion, H., Lam, W., Blitzer, J., Çelebi, A., Qi, H., Drabek, E. y Liu, D. 2002, *Evaluation of text summarization in a cross-lingual information retrieval framework*, [Online], Informe Técnico, Johns Hopkins University, Disponible en: <http://www.clsp.jhu.edu/ws2001/groups/asmd/jhu01summ-finalreport.ps> [30 Junio 2005]

Rapp, R. 1999, "Automatic identification of word translations from unrelated English and German corpora", en *Proc. of the 37th Annual Meeting of the ACL*, pp. 1-17.

Rauber, A. y Merkl, D. 1999, "The SOMLib Digital Library System", en *European Conference on Digital Libraries*, pp. 323-342.

Reimer, U. y Hahn, U. 1988, "Text condensation as knowledge base abstraction", en *Proceedings of the Fourth Conference on Artificial Intelligence Applications*, pp. 338–344.

Real Academia Española, 2001, *Diccionario de la Lengua Española*.

Rennie, J. y Rifkin, R. 2001, *Improving multiclass text classification with the support vector machine*, informe técnico, MIT.

Riloff, E. 1995, "Little words can make a big difference for text classification", en *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 130-136.

Robertson, S.E. 2004, "Understanding inverse document frequency: on theoretical arguments for IDF", *Journal of Documentation*, vol. 60, pp. 503-520.

Robertson, S.E. y Spärck-Jones, K. 1976, "Relevance weighting of search terms", *Journal of the ASIS*, vol. 27, no. 3, pp. 129-146.

Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M. y Gatford, M. 1994, "Okapi at TREC-2", en *Text REtrieval Conference*, pp. 21-34.

Rocchio, J.J. 1966, *Document retrieval systems - optimization and evaluation*, Disertación doctoral, Harvard University.

Rocchio, J.J. 1971, "Relevance Feedback in Information Retrieval", en Salton, G. (Ed.): *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313-323. Prentice-Hall.

Rogati, M. y Yang, Y. 2001, "Cross-Lingual Pseudo-Relevance Feedback using a Comparable Corpus: CLEF Working Notes", *CLEF 2001*, pp. 151-157.

Rosenblatt, F. 1958, "The Perceptron: A probabilistic model for information storage and organization in the brain", *Psychological Review*, 65, pp. 386-408.

Roussinov, D. y Chen, H. 1998, "A scalable self-organizing map algorithm for textual classification: A neural network approach to thesaurus generation", *Communication and Cognition*, vol. 15, no. 1-2, pp. 81-112.

Rubner, Y., Tomasi, C. y Guibas, L.J. 2000, "The Earth Mover's Distance as a Metric for Image Retrieval", *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99-121.

Rucker, J. y Marcos, J.P. 1997, "Siteseer: Personalized Navigation for the Web", *Communications of the ACM*, vol. 40, no. 3, pp. 73-75.

Rudman, J. 1998, "The State of Authorship Attribution Studies: SomeProblems and Solutions", *Computers and the Humanities*, 31, pp. 351-365.

Ruiz, M. E. y Srinivasan, P. 1997, "Automatic Text Categorization Using Neural Networks", en *Proceedings of the 8th ASIS SIG/CR Classification Research Workshop*, pp. 59-72.

Ruiz, M.E. y Srinivasan, P. 1999, "Hierarchical neural networks for text categorization", en *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 281-282.

Saggion, H. y Gaizauskas, R. 2004, "Multi-Document Summarization by Cluster/Profile Relevance and Redundancy Removal", *DUC 2004*.

Sahami, M., Dumais, S., Heckerman, D. y Horvitz, E. 1998, "A bayesian approach to filtering junk e-mail", en *AAAI-98 Workshop on Learning for Text Categorization*.

Sahr, R. 2004, *Inflation Conversion Factors for Dollars 1665 to Estimated 2014*, [Online], Oregon State University, Disponible en: <http://oregonstate.edu/dept/pol_sci/fac/sahr/sahr.htm> [30 Junio 2005].

Salton, G. 1968, *Automatic Content Analysis in Information Retrieval*, Informe técnico, Cornell University.

Salton, G. 1968, *Automatic Information Organisation and Retrieval,* McGraw-Hill. Citado en Jardine, N. y van Rijsbergen, C.J. 1971, "The use of hierarchic clustering in information retrieval", *Information Storage and Retrieval*, vol. 7, pp. 217-240.

Salton, G. 1972, *Automatic processing of current afairs queries*, Informe Técnico, Cornell University.

Salton, G. y Crouch, D.B. 1989, *User-System Interaction in Automatic Information Retrieval*, Informe Técnico, Cornell University.

Salton, G. y Lesk, M.E. 1965, "The SMART Automatic Document Retrieval System – An Illustration", *Communications of the ACM*, vol. 8, no. 6, pp. 391-398.

Salton, G. y Singhal, A. 1994, *Automatic text theme generation and the analysis of text structure*, Informe Técnico, Cornell University.

Salton, G., Fox, E.A. y Wu, H. 1983, "Extended Boolean information retrieval", *Communications of the ACM*, vol. 26, no. 11, pp. 1022-1036.

Salton, G., Singhal, A., Buckley, C. and Mitra, M. 1996, "Automatic Text Decomposition Using Text Segments and Text Themes", en *Proceedings of the Seventh ACM Conference on Hypertext*, pp.53-65.

Salton, G., Singhal, A., Mitra, M. y Buckley, C. 1997, "Automatic Text Structuring and Summarization", *Information Processing and Management*, vol. 33, no. 2, pp. 193-207.

Santos Jr., E., Mohamed, A.A. y Zhao, Q. 2004, "Automatic Evaluation of Summaries Using Document Graphs", en *Proceedings of the ACL-04 Workshop*, pp. 66-73.

Schapire, R.E. 1990, "The strength of weak learnability", *Machine Learning*, vol. 5, no. 2, pp. 197-227.

Schapire, R.E. y Singer, Y. 2000, "BoosTexter: a boosting-based system for text categorization", *Machine Learning*, vol. 39, no. 2-3, pp. 135-168.

Schapire, R.E., Singer, Y. y Singhal, A. 1998, "Boosting and Rocchio applied to text filtering", en *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pp. 215-223.

Schohn, G. y Cohn, D. 2000, "Less is more: Active learning with support vector machines", en *Proceedings of the 17th International Conference on Machine Learning*, pp. 839-846.

Schütze, H., Hull, D.A. y Pedersen, J.O. 1995, "A comparison of classifiers and document representations for the routing problem", en *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, pp. 229-237.

Seaborne, A. 2004, *RDQL - A Query Language for RDF*, [Online], Propuesta al W3C, Hewlett-Packard, Disponible en: <http://www.w3.org/Submission/RDQL/> [30 Junio 2005].

Sebastiani, F. 2002, "Machine learning in automated text categorization", *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47.

Selberg, E. y Etzioni, O. 1995, "Multi-Service Search and Comparison Using the MetaCrawler", en *Proceedings of the 4th World Wide Web Conference*.

Shardanand, U. y Maes, P. 1995, "Social Information Filtering: Algorithms for Automating «Word of Mouth»", en *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, pp. 210-217.

Sibun, P. y Reynar, J.C. 1996, "Language Identification: Examining the Issues", en *5th Symposium on Document Analysis and Information Retrieval*, pp. 125-135.

Silverstein, C., Henzinger, M., Marais, H. y Moricz, M. 1998, *Analysis of a Very Large AltaVista Query Log*, Informe técnico, Digital Systems Research Center.

Singhal, A, Mitra, M. y Buckley, C. 1997, "Learning routing queries in a query zone", en *Proceedings of SIGIR'97*, pp. 25-32.

Singhal, A., Buckley, C. y Mitra, M. 1996, "Pivoted Document Length Normalization", en *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 21-29.

Smalheiser, N.R. y Swanson, D.R. 1998, "Using Arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses", *Computer Methods and Programs in Biomedicine*, vol. 57, pp. 149-153.

Sneath, P.H.A. y Sokal, R.R. 1973, *Numerical Taxonomy*, Freeman, Londres.

Soergel, D. 1999. "The Rise of Ontologies or the Reinvention of Classification", *Journal of the American Society for Information Science*, vol. 50, no. 12, pp. 1119-1120.

Spärck-Jones, K. 1972, "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, vol. 28, no. 1, pp. 11-21.

Spärck-Jones, K. 1974, "Automatic indexing", *Journal of Documentation*, vol. 30, no. 4, pp. 393-432. Citado en Rasmussen, E. 2002, "Evaluation in Information Retrieval", en *Proceedings of the ISMIR 2002*, pp. 45-49.

Spärck-Jones, K. 1979, "Experiments in relevance weighting of search terms", *Information Processing and Management*, vol. 15, no. 3, pp. 133-144.

Spärck-Jones, K. 1981, *Information Retrieval Experiment*, Butterworth-Heinemann, Reino Unido.

Spärck-Jones, K. 1999, "Automatic summarizing: Factors and directions", en Mani, I. y Maybury, M. (Eds.), *Advances in automatic text summarization*, pp. 2-12.

Spärck-Jones, K. y van Rijsbergen, C.J. 1975, *Report on the Need for and Provision of an "Ideal" Information Retrieval Test Collection*, Informe técnico, University of Cambridge.

Spärck-Jones, K. y Webster, C.A. 1979, *Research in Relevance Weighting*, Informe técnico, University of Cambridge.

Spärck-Jones, K., van Halteren, H., Moens, M-F., Lapalme, G., Radev, D., Dorr, B., Over, P., Hovy, E., McKeown, K. y Harman, D. 2004, *Road Map for DUC 05-DUC 07 Automatic Summarising Evaluation Programme*, [Online], Disponible en: <http://www-nlpir.nist.gov/projects/duc/RM0507/rm.html> [30 Junio 2005].

Speyer, B. y Allen, W. 1994, *New Directory Services*, [Online], Google Inc., Disponible en: <http://groups.google.com/groups?selm=9401211534.AA26997%40faith.mcc.com&output=gplain> [30 Junio 2005].

Staab, S., Erdmann, M., Mädche, A. y Decker, S. 2000, "An extensible approach for Modeling Ontologies in RDF(S)", en *Proceedings of the ECDL-2000 Workshop "Semantic Web: Models, Architectures and Management"*.

Stamatos, E., Fakotakis, N. y Kokkinakis, G. 2001, "Computer-based Authorship Attribution without lexical Measures", [Online], *Computers and the Humanities*, vol. 35, no. 2, pp. 193-214, disponible en: <http://slt.wcl.ee.upatras.gr/papers/stamatatos6.pdf> [30 Junio 2005]

Steinbach, M., Karypis, G. y Kumar, V. 2000, *A Comparison of Document Clustering Techniques*, Informe Técnico, Universidad de Minnesota.

Steinberg, S.G. 1996, "Seek and Ye Shall Find (Maybe)", *Wired*, [Online], vol. 4, no. 5, Disponible en: <http://www.wired.com/wired/archive/4.05/indexweb.html> [30 Junio 2005].

Stevens, M.E. 1970, *Automatic Indexing: A state-of-the-art report*, US National Bureau of Standards, Washington.

Stiles, H.E. 1961, "The association factor in information retrieval", *Journal of the ACM*, vol. 8, pp. 271-279 (citado por van Rijsbergen, C.J. 1979, *Information Retrieval, 2nd Edition*. Butterworth-Heinemann, EE.UU.)

Swadesh, M. 1950, "Salish internal relationships", *International Journal of American Linguistics*, 16, pp. 157-167.

Swanson, D.R. 1977, "Information retrieval as a trial-and-error process", *Library Quarterly*, vol. 47, no. 2.

Swanson, D.R. 1986, "Undiscovered public knowledge", *Library Quarterly*, vol. 56, no. 2, pp. 103-118.

Swanson, D.R. 1991, "Complementary structures in disjoint science literatures", en Bookstein, A., Chiaramella, Y., Salton, G. y Raghavan, V.V. (Eds.): *SIGIR'91*, pp. 280-289.

Swanson, D.R. y Smalheiser, N.R. 1997, "An interactive system for finding complementary literatures: A stimulus to scientific discovery", *Artificial Intelligence*, vol. 91, no. 2, pp. 183-203.

Tait, J.I. 1983, *Automatic Summarizing of English Texts*, disertación doctoral, University of Cambridge.

Tombros, A., Villa, R. y van Rijsbergen, C.J. 2002, "The effectiveness of query-specific hierarchic clustering in information retrieval", *Information Processing and Management: an International Journal*, vol. 38, no. 4, pp. 559-582.

Tong, S. y Koller, D. 2000, "Support vector machine active learning with applications to text classification", en *Proceedings of the 17th International Conference on Machine Learning*, pp. 999- 1006.

U.S. Department of Commerce, 2002. *A Nation Online: How Americans are Expanding Their Use of the Internet*, Washington, D.C.

Ursing, B.M. y Arnason, U. 1998, "Analyses of mitochondrial genomes strongly support a hippopotamus-whale clade", *Proceedings of the Royal Society of London*. Series B, Biological Sciences, vol. 265, pp. 2251-2255.

Valiant, L. 1984, "A theory of the learnable", *Communications of the ACM*, vol. 27, no. 11, pp. 1134-1142.

van Halteren, H. 2002, "Writing Style Recognition and Sentence Extraction", en *DUC'02 Conference Proceedings*.

van Harmelen, F., Patel-Schneider, P.F. y Horrocks, I. 2001, *Reference Description of the DAML+OIL (March 2001) Ontology Markup Language*, Informe técnico, Disponible en: <http://www.daml.org/2001/03/reference.html> [30 Junio 2005].

van Rijsbergen, C.J. 1977, "A theoretical basis for the use of co-occurrence data in information retrieval", *Journal of Documentation*, vol. 33, pp. 106-119.

van Rijsbergen, C.J. 1979, *Information Retrieval, 2nd Edition*. Butterworth-Heinemann, EE.UU.

Vanderwende, L., Banko, M. y Menezes, A. 2004, "Event-centric summary generation", en *Document Understanding Conference at HLT-NAACL*.

Varré, J.S., Delahaye, J.P. y Rivals, É. 1999, "The Transformation Distance: A Dissimilarity Measure Based on Movements of Segments", *Bioinformatics*, vol. 15, no. 3, pp 194-202.

Verdaguer, P. 1999, *Grammaire de la langue catalane. Les origines de la langue*, Curial.

Wang, L., Fan, W., Yang, R., Xi, W., Luo, M., Zhou, Y. y Fox, E.A. 2004, "Ranking Function Discovery by Genetic Programming for Robust Retrieval", en *Proceedings of the Twelfth TREC Conference*.

Warnow, T., Evans, S.N., Ringe, D. y Nakleh, L. 2004, *Stochastic models of language evolution and an application to the Indo-European family of languages*, Informe técnico, Universidad de California en Berkeley.

Watanabe, Y., Murata, M., Takeuchi, M. y Nagao, M. 1996, "Document Classification Using Domain Specific Kanji Characters Extracted by χ^2 Method", en *Proceedings of the 16th COLING*, pp. 794-799.

Weigend, A.S., Wiener, E.D. y Pedersen, J.O. 1999, "Exploiting Hierarchy in Text Categorization", *Information Retrieval*, vol. 1, no. 3, pp. 193-216.

Weiss, S.M., Apté, C., Damerau, F.J., Johnson, D.E., Oles, F.J., Goetz, T. y Hampp, T. 1999, "Maximizing text-mining performance", *IEEE Intelligent Systems*, vol. 14, no. 4, pp. 63-69.

Whitworth, B. y Whitworth, E. 2004, "Spam and the Social-Technical Gap", [Online], *Computer*, vol. 37, no. 10, pp. 38-45, disponible en: <http://www.computer.org/computer/homepage/1004/whitworth/rx038.pdf> [30 Junio 2005]

Wiener, E.D., Pedersen, J.O. y Weigend, A.S. 1995, "A neural network approach to topic spotting", en *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, pp. 317-332.

Wong, S.K.M., Ziarko, W. y Wong, P.C.N. 1985, "Generalized Vector Space Model in Information Retrieval", en *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 18-25.

Yang, Y. y Chute, C.G. 1994, "An example-based mapping method for text categorization and retrieval", *ACM Transactions on Information Systems*, vol. 12, no. 3. Citado por Han, E., Karypis, G. y Kumar, V. 1999, *Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification*, Informe técnico, Universidad de Minnesota.

Yang, Y. y Liu, X. 1999, "A re-examination of text categorization methods", en *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, pp. 42-49.

Zechner, K. 1996, "Fast generation of abstracts from general domain text corpora by extracting relevant sentences", en *Proceedings of COLING-96*, pp. 986-989.

Zhang, L. y Yao, T. 2003, "Filtering Junk Mail with A Maximum Entropy Model", [Online], en *Proceedings of 20th International Conference on Computer Processing of Oriental Languages (ICCPOL03)*, disponible en: <http://homepages.inf.ed.ac.uk/s0450736/paper/junk.pdf> [30 Junio 2005]

Zhang, L., Zhu, J. y Yao, T. 2004, "An evaluation of statistical spam filtering techniques", *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 3, no. 4, pp. 243-269.

Zhao, Y. y Karypis, G. 2002, *Criterion Functions for Document Clustering: Experiments and Analysis*, Informe técnico, Universidad de Minnesota.