Los modelos clásicos de Recuperación de información y su vigencia

Juan Antonio Martínez Comeche
Departamento de Biblioteconomía y Documentación
Universidad Complutense de Madrid
comeche@ccdoc.ucm.es

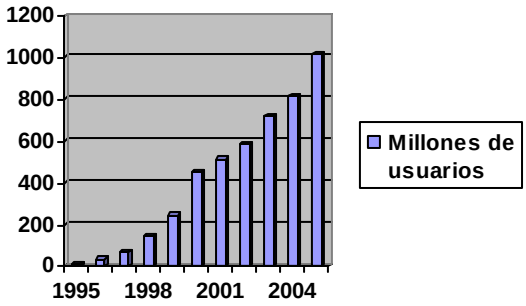
Introducción

Aunque desde mediados del siglo XX se viene trabajando en el área de la Recuperación de información, en los últimos diez años su relevancia ha aumentado notablemente. Entre otros posibles factores desencadenantes de este efecto, quisiera destacar dos: en primer lugar, el crecimiento espectacular y constante de la web, con el consiguiente aumento en el número de documentos digitales a disposición de los usuarios de la red. En segundo lugar, el cambio producido en los hábitos de los usuarios a raíz de la preponderancia de internet entre las diversas modalidades de acceso a la información, lo que ha traído consigo una modificación paralela en los servicios que demanda.

Desde que Berners-Lee inventó la World Wide Web en 1989 mientras trabajaba en la Centro Europeo para la Investigación Nuclear (CERN, actualmente Organisation Européenne pour la Recherche Nucléaire)[1], el número de usuarios de la red ha sufrido una evolución imparable. Inicialmente fueron 50 personas las que en 1989 compartían páginas web[2], pero solo cinco años más tarde se estimaba en 16 millones el número de usuarios en todo el mundo. Al cabo de otros cinco años, en 2000, la cifra de usuarios asciende a 451 millones, y a finales de 2005 se alcanzan mil millones de usuarios[3]. Se puede observar el resumen de este crecimiento espectacular en la gráfica 1.

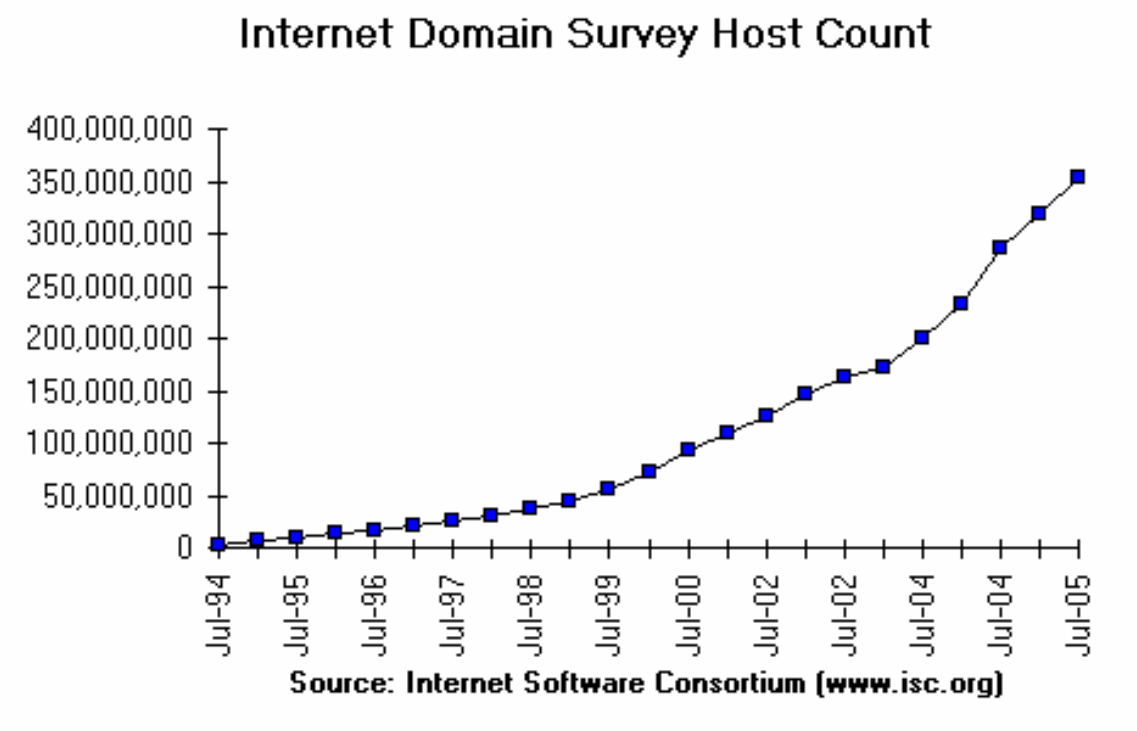
[1] Vid. Los orígenes de internet en http://www.w3.org/History.html. [Consulta: 16/03/2006]
[2] Gil, P. How Big is the Internet? (2005). Disponible en http://netforbeginners.about.com/cs/technoglossary/f/FAQ3.htm. [Consulta: 16/03/2006]
[3] Datos tomados de Internet World Stats (2006). Internet growth statistics. Disponible en http://www.internetworldstats.com/emarketing.htm. [Consulta: 15/03/2006]

Grafica 1



Este desarrollo ha provocado que el número de documentos disponible en la red haya sufrido una evolución semejante, a juzgar por el crecimiento en la cantidad de servidores en funcionamiento en la red, conforme se resume en la siguiente gráfica[4].

Gráfica 2



[4] Internet Systems Consortium (2006). Internet Domain Survey Host Count. Disponible en http://www.isc.org/index.pl?/ops/ds/. [Consulta: 16/03/2006]

De hecho, estudios recientes cifran en al menos 11’5 mil millones el número de páginas web disponibles a principios de 2005[5]. Como es fácil de comprender con semejantes cifras, resulta imposible pensar siquiera en catalogar o clasificar manualmente esta gigantesca amalgama de documentos. En consecuencia, la Recuperación de información, entendida como el área de conocimiento a la que atañe la representación, el almacenamiento, el tratamiento y el acceso automatizados a los documentos o a sus sustitutos[6], surge como la única vía posible para tratar de controlar este volumen ingente de información digital.

En cuanto al segundo factor enunciado al comienzo, hemos visto cómo actualmente miles de millones de usuarios emplean habitualmente la web como medio para acceder y consultar aquella información que precisan. Ante semejante volumen de documentación, los buscadores y metabuscadores se han convertido en una herramienta imprescindible para discernir las poquísimas páginas que incluyen la información buscada frente a los millones de páginas restantes que resultan irrelevantes. De hecho, a finales de 2006, el empleo de un buscador para localizar información se convertía en la segunda actividad más frecuente realizada por los usuarios de Internet en Estados Unidos, empatada ya a la actividad tradicionalmente más popular, esto es, el envío o lectura de correo electrónico[7]. Acostumbrados cada vez más al empleo de las nuevas tecnologías de la comunicación y a las técnicas de recuperación mediante la formulación de consultas (introduciendo habitualmente una o varias palabras del lenguaje natural), los usuarios demandan nuevos productos y servicios en el ámbito bibliotecario, desde el acceso en línea a la biblioteca –que deja de estar abierta en un horario limitado- hasta la posibilidad de acceder a la información mediante otras herramientas distintas de la clásica lectura del catálogo –consulta de fondos mediante un módulo de recuperación semejante al de los buscadores, navegación por el fondo mediante interfaces gráficos- o a otro tipo de información más allá del propio catálogo –páginas web de otras bibliotecas u otras unidades de información, páginas web de proveedores de información (editores de revistas, por ejemplo), directorios, bases de

[5] Gulli, A.; Signorini, A. (2006). The indexable web is more than 11.5 billion pages. Disponible en http://www.cs.uiowa.edu/~asignori/web-size/. [Consulta: 16/03/2006]

[6] Definición basada en la de Salton, G.; McGill, M.J. Introduction to modern Information Retrieval. New York: Mc.Graw-Hill, 1983, p. 7. Se le ha añadido el proceso de tratamiento (pensando principalmente en la indización automática) y el carácter automatizado en todos ellos.

[7] Pew Internet & American Life Project Tracking surveys (March 2000-December 2006). Disponible en http://www.pewinternet.org/trends/Internet_Activities_1.11.07.htm [Consulta: 05/05/2007]

datos, libros electrónicos, etc.-. Este cambio se percibe en los conceptos de biblioteca digital o biblioteca virtual a los que se tiende cada vez en mayor medida "a raíz de la creciente demanda por […] un acceso coherente a bases de datos extensas y geográficamente dispersas"[8].

En consecuencia, la Recuperación de información se ha ido convirtiendo en un campo de conocimiento cada vez más necesario al que acudir en busca de soluciones automatizadas no solo cuando hablamos de búsqueda de información en internet, sino en el propio ámbito bibliotecario al facilitar la creación de productos y servicios acordes con las nuevas demandas de los usuarios.

Son muchos los enfoques que se han experimentado para abordar el objetivo esencial de la Recuperación de información (RI), esto es, la recuperación de todos los documentos relevantes y al mismo tiempo rechazar todos los documentos irrelevantes ante la formulación de una consulta por parte del usuario: desde el modelo booleano (por tratarse de una de las vías más simples desde el punto de vista teórico) hasta la aplicación de técnicas de Inteligencia artificial, entre las que podemos destacar las redes neuronales (RN), los algoritmos genéticos (AG) o el Procesamiento del lenguaje natural (PLN)[9]. En este trabajo ahondaremos en los principios teóricos de los denominados modelos clásicos de RI: el booleano, el probabilístico y el vectorial, comentando al tiempo su vigencia en los sistemas de recuperación de información (SRI) actuales.

Modelo booleano

Constituye el primer modelo teórico, el más antiguo, empleado para establecer el subconjunto de documentos relevantes, en relación a una consulta específica, de entre todos los que configuran la colección (ya se trate del fondo de una biblioteca o de todas las páginas disponibles en la web). Al mismo tiempo es, sin duda, uno de los más sencillos tanto desde un punto de vista teórico como práctico, al basarse en la teoría de

[8] DIGITAL LIBRARIES INITIATIVE (2006). Disponible en http://dli.grainger.uiuc.edu/national/spanish/index.html. [Consulta: 16/03/2006].
[9] Ellis, D. Progress and problems in information retrieval. London: LA, 1996.

conjuntos y en el álgebra de Boole –por una parte- y al ser fácil de diseñar e implementar en la práctica, por otra parte[10].

El procesamiento automatizado de un documento textual comienza con la extracción de los términos de indización, es decir, los términos que van a ser utilizados para describir el contenido del documento. La posibilidad más simple consiste en considerar todas las palabras aisladas que aparecen en el texto como los términos de indización. Habitualmente se eliminan algunas –las denominadas palabras vacías, entre las que suelen figurar los números, las preposiciones, conjunciones, verbos ser, haber y estar-, aunque la consideración o no de estos procesos añadidos (como la inclusión de una lista de palabras vacías) no influyen en absoluto sobre los principios teóricos del modelo. Una vez extraídas las palabras del texto, se ordenan por orden alfabético y se guardan en un denominado **fichero inverso**[11], junto con la referencia del documento de donde proceden (normalmente un número de documento asignado previamente por el sistema). Si se repite este proceso con todos los documentos de la colección, obtendremos finalmente un fichero inverso que almacena los siguientes datos:

- En primer lugar, los términos de indización (las palabras) que aparecen en toda la colección (ya sean los propios textos, los resúmenes de los textos del fondo y/o los títulos).
- En segundo lugar, cada uno de dichos términos (palabras) incorpora una lista con los números de los documentos en los que aparece.

Conviene destacar en este proceso que no se ha guardado noticia alguna sobre la frecuencia de aparición de cada término en cada documento. De ahí que el modelo booleano clásico sea denominado **modelo binario**, pues de la consulta del fichero inverso únicamente puedo saber si un determinado término de indización está presente

[10] Sobre el modelo booleano pueden consultarse Baeza-Yates, R.; Ribeiro-Neto, B. Modern information retrieval. New York: ACM, 1999; Korfhage, R. Information storage and retrieval. New York: John Wiley & Sons, 1997; Rijsbergen, C. J. van. Information retrieval. London: Butterworths, 1979. Disponible en http://www.dcs.gla.ac.uk/Keith/Preface.html [Consulta: 16/03/2006].

[11] El fichero inverso consta, en realidad, de varios ficheros. Así, la relación de términos empleados en la representación de los documentos de la colección se guarda en el llamado **fichero diccionario**. Se simplifican estos detalles con el ánimo de destacar lo esencial en el modelo. Sobre el fichero inverso vid. Moya Anegón, F. de. Los sistemas integrados de gestión bibliotecaria: estructuras de datos y recuperación de información. Madrid: ANABAD, 1995.

(en cuyo caso se simbolizará por el número 1) o está ausente (en cuyo caso se simbolizará por el número 0) en cada uno de los documentos de la colección.

De manera que el fichero inverso (en concreto el fichero diccionario) puede representarse por una tabla cuyos datos básicos son los siguientes:

Tabla 1. Datos mínimos del fichero inverso

	D1	D2	…..	Dn
T1	1	0	…..	1
T2	0	1	…..	0
…..	…..	…..	…..	…..
Tt	0	1	…..	1

donde T1, T2, …, Tn son los términos de indización empleados en la colección; D1, D2, …, Dn son los documentos que componen la colección; y donde el "1" significa que el término correspondiente aparece en ese documento concreto, mientras que el "0" significa que el término no aparece en dicho documento. Ello implica que no se tiene en cuenta la frecuencia de aparición de los términos en los documentos: tanto si aparece veinte veces como si aparece una sola vez, en todos los casos ese término en dicho documento se representará mediante un "1", reservándose el "0" para cuando no aparezca. Se comprende entonces la denominación de **modelo binario** que recibe, pues únicamente se juega con dos posibilidades: la aparición y la no aparición de los descriptores en los documentos.

Si observamos la tabla, podemos deducir de ella las dos representaciones empleadas al manejar el modelo binario. Por una parte, cada término de indización se representa por la lista de documentos en los que aparece, lo que implica la observación de la tabla por filas:

T1 = {D1, …., Dn}

T2 = {D2, …..}

…….

Tt = {D2, …., Dn}

Por otra parte, cada documento se representa por la lista de ceros y unos correspondientes a los términos de indización que contiene, lo que implica la observación de la tabla por columnas:

D1 = {1, 0, ….. , 0}
D2 = {0, 1, …..., 1}
……..
Dn = {1, 0, ……, 1}

Se comprende bien ahora por qué se dice que en el modelo binario todo documento se representa mediante una serie ordenada de ceros y unos, tantos como términos de descripción se empleen en la colección: el primer número corresponderá siempre a T1, el segundo dígito corresponderá a T2, y así sucesivamente hasta llegar a Tt, siendo t el número de descriptores distintos que representan el contenido de esa colección.

La misma tabla nos servirá para explicar el método empleado por los SRI basados en este modelo para contestar a las consultas formuladas por los usuarios. En primer lugar, el usuario debe introducir palabras, precisamente aquéllas que describan su necesidad informativa, o una fórmula que se ajuste a la sintaxis booleana. Habitualmente los usuarios empleamos pocas palabras, de manera que muchos SRI presentan un número máximo de palabras posibles en la consulta que ronda la decena (Google, por ejemplo). Si el usuario se limita a introducir dos palabras, por ejemplo, el sistema automáticamente convertirá dicha consulta a una fórmula booleana, introduciendo entre las dos palabras una conectiva por defecto, habitualmente AND.

Con un ejemplo se comprenderá fácilmente el procedimiento seguido por un SRI. Supongamos que deseamos localizar documentos sobre las plantaciones de café en Colombia. Lo que solemos hacer es introducir en la ventana de un buscador en Internet esa frase tal cual. El SRI analiza la cadena de caracteres introducida en la ventana y la trata en principio como si se tratase de un documento más, aunque en ocasiones puede ser sometida a tratamientos específicos (análisis sintáctico débil, por ejemplo, imposible si se tratase de un documento largo). En consecuencia, eliminará las palabras vacías (imaginemos que “de” y “en” lo son, lo que no resultaría extraño en español), resultando

las palabras **plantaciones**, **café** y **Colombia**. A continuación, el SRI introduce entre ellas la conectiva por defecto, normalmente AND como ya dijimos. La fórmula en este caso quedaría:

plantaciones AND **café** AND **Colombia**

A continuación el sistema trata de localizar la palabra en su fichero inverso. Pueden suceder dos cosas: que figure o que no figure entre los términos de indización almacenados en el SRI:

* Si figura, sustituye la palabra por el conjunto/lista de documentos de la colección donde aparecen dicha palabra/término de indización.

* Si no figura, lo que suele ser muy extraño en SRI cuyas colecciones abarcan millones de documentos de todo tipo, sustituye dicha palabra por el conjunto vacío (no aparece en ningún documento de la colección). De ahí que cuando introducimos una sola palabra y ésta no aparece en el fichero inverso, y por tanto no se halla en ninguno de los documentos de la colección, un SRI inspirado en el modelo booleano muestre la siguiente respuesta: “No existe en la colección ningún documento que incluya dichas palabras”. De igual forma, si introducimos una palabra muy poco frecuente (como “supercalifragilist”, por ejemplo) junto a otra muy frecuente (“caballos”, por ejemplo), considerando que la conectiva por defecto es AND, el SRI únicamente mostrará entre los resultados los escasísimos documentos (quizá 2, 3 ó 4) en los cuales aparecen simultáneamente ambos términos. Es lo que sucede con los buscadores habituales de Internet, por ejemplo, prueba de que están basados en el modelo booleano, aunque ciertamente no de manera exclusiva, pues todos ellos en la actualidad añaden o superponen características de otros modelos, habitualmente el vectorial, como veremos más adelante, para mejorar los resultados que ofrecerán al usuario.

Por último, con las listas que sustituyen a las palabras el sistema efectúa las operaciones de conjuntos correspondientes a las conectivas que figuren en la consulta, de la siguiente manera:

- El resultado de dos listas/conjuntos de documentos unidos por la conectiva AND da como resultado el conjunto de los documentos en los que aparecen simultáneamente ambos términos.
- El resultado de dos listas/conjuntos de documentos unidos por la conectiva OR da como resultado el conjunto de los documentos que en los que aparece

el primer término (y no el segundo), o el segundo término (y no el primero) o ambos términos simultáneamente.

- El resultado de una lista/conjunto de documentos precedido por la conectiva NOT da como resultado el conjunto de los documentos de la colección en los que no aparece el término.

En nuestro ejemplo, tras las operaciones de conjuntos correspondientes a la consulta: “plantaciones AND café AND Colombia”, obtendríamos el conjunto de los documentos de la colección en los que aparecen simultáneamente las palabras plantaciones, café y Colombia.

Una propiedad importante de los sistemas de recuperación basados en el modelo booleano es que no pueden efectuar ningún proceso de ordenación con los documentos resultantes de la búsqueda, pues todos ellos cumplen la fórmula en idénticas condiciones. Esta característica suele denominarse **equiparación exacta**, impidiendo que el sistema pueda situar en primer lugar aquel documento posiblemente más útil o relevante para el usuario y relegando a las últimas posiciones a aquellos otros documentos con menos probabilidades de ser relevantes en relación a la consulta. ¿Cómo podría efectuar el sistema una ordenación con los documentos de la respuesta? Existen muchas posibilidades, pero una muy sencilla que permite comprender el proceso de clasificación consistiría en ordenar los documentos por el número total de veces que aparece alguna de las palabras en ellos. Así, el primer documento podría contabilizar 45 apariciones (10 veces aparece “plantaciones”, “café” surge 30 veces y “Colombia” aparece 5 veces, por ejemplo), el segundo documento contabilizaría 31 apariciones, el tercero incluiría 12 apariciones, y así sucesivamente en orden decreciente. Dado que los buscadores en Internet llevan a cabo un proceso de ordenación de los resultados, tratando de situar primeramente los más relevantes en función de la consulta del usuario y relegando los menos útiles, deducimos que dichos buscadores no se inspiran únicamente en el modelo booleano, sino que efectivamente combinan o añaden características de otros modelos.

Como podemos observar, el carácter binario (consideración exclusivamente de la presencia/ausencia de los términos en los documentos) es el principal responsable de la equiparación exacta, siendo considerado la principal desventaja del modelo. De hecho,

la ponderación se ha demostrado muy útil para mejorar los resultados de la recuperación.

Como esta desventaja puede superarse mediante la superposición o inclusión de características de otros modelos, esencialmente el vectorial, no debe extrañarnos que todavía hoy siga siendo un modelo presente –aunque no de manera exclusiva, insisto, y perfeccionado con la combinación de otros modelos- en los sistemas de recuperación de información. Muchos motores de búsqueda en la web utilizan en un estadio inicial este modelo, al que se superponen otros modelos antes de dar la respuesta a una determinada consulta, por ser de desarrollo sencillo (como hemos visto, en su versión básica solamente involucra el empleo de un fichero inverso y una interfaz de consulta que permita computar consultas expresadas mediante palabras o expresiones booleanas), fácil de utilizar por parte de un usuario medio (basta introducir palabras relativas a la necesidad informativa), y bastante eficaz en los resultados obtenidos (en gran parte debido al volumen ingente de documentación presente en la red, lo que provoca que la reducción de la respuesta a los documentos que satisfagan estrictamente las condiciones de la consulta –por defecto, recordemos, la conectiva AND- aún genera subconjuntos muy abultados de documentos).

Modelo probabilístico

Introducido en la década de los setenta por Robertson y Sparck Jones, también es conocido como modelo de recuperación de independencia binaria (BIR)[12]. Este modelo se basa en las siguientes consideraciones, pues evitaremos el empleo de fórmulas matemáticas, incidiendo en las ideas que subyacen al modelo:

- Para caracterizar los documentos de la colección se han empleado ciertos términos de indización (palabras en principio).

[12] Sobre el modelo probabilístico pueden consultarse, además de los textos de carácter general citados a propósito del modelo booleano, los artículos de Robertson, S. E. The probability ranking principle in IR. Journal of Documentation, 1977, 33(4):294-304; Sparck Jones, K. Search term relevance weighting given little relevante information. Journal of Documentation, 1979, 35(1): 30-48; Croft, W. B.; Harper, D. J. Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 1979, 35(4):285-295. Los tres últimos pueden consultarse también en READINGS in information retrieval. (eds.) Sparck Jones, K.; Willett, P. San Francisco: Morgan Kaufmann, 1997.

- Dada una necesidad informativa del usuario, existe un subconjunto de documentos de la colección que contiene exclusivamente los documentos relevantes en relación a ella.
- Si el usuario supiese los términos de indización que permiten caracterizar tal subconjunto de documentos relevantes (porque aparecen en ellos y no aparecen en el resto de los documentos de la colección), tendríamos el problema resuelto. Como vemos, el modelo probabilístico parte exclusivamente de la presencia o ausencia de los términos en los documentos de la colección. Se trata, pues, también de un **modelo binario**, como el modelo booleano.
- Lamentablemente, en un caso real el usuario no sabe cuáles son los términos de indización que configurarían la **consulta ideal**. Tampoco sabe, de hecho, en qué medida los términos empleados en la consulta permiten discernir los documentos relevantes y rechazar simultáneamente los documentos irrelevantes.
- El modelo probabilístico actúa precisamente sobre los términos que configuran la consulta del usuario, ponderándolos; esto es, imponiéndoles un peso o número a cada uno de ellos, mayor cuanto mejor permita discernir los documentos relevantes de los irrelevantes, y menor en caso contrario. De esta manera se persigue que el sistema efectúe la recuperación incidiendo sobre todo en los mejores descriptores de entre los empleados por el usuario en la consulta, minimizando la importancia de aquellos otros términos que, aun figurando en la consulta, son malos descriptores del conjunto respuesta ideal.
- Como tampoco se puede saber a priori cuáles, de entre los términos que configuran la consulta, son buenos descriptores y cuáles no lo son, a este modelo no le queda otro remedio que considerar, para cada uno de los términos empleados en la consulta, la “probabilidad de ser buen descriptor” (probabilidad de que el término empleado en la consulta esté presente en un documento del conjunto de documentos relevantes en relación a la consulta) y simultáneamente, para ese mismo término, la “probabilidad de ser mal descriptor” (probabilidad de que ese mismo término esté presente en un

documento del conjunto de documentos irrelevantes en relación a la consulta).

- Ahora bien, como estas probabilidades –insistimos, para cada uno de los términos empleados en la consulta- son desconocidas en el momento de formalizar dicha consulta, este modelo se ve en la necesidad de efectuar una **hipótesis inicial** sobre sus valores. La obligatoriedad de hacer una hipótesis inicial sobre las "probabilidades de ser buen y mal descriptor" para cada término de la consulta se ha considerado el principal inconveniente de este modelo. De ahí que siga siendo una de las áreas de investigación más activas entre sus especialistas.

Basados en estos pesos iniciales, el modelo probabilístico es capaz de calcular el grado de similitud existente entre cada documento de la colección y la consulta ponderada, consiguiendo ordenar los documentos de la colección en orden descendente de probabilidad de relevancia en relación a la consulta. De esta manera el modelo probabilístico supera el gran inconveniente puesto de manifiesto en el modelo booleano, a saber, la equiparación exacta. En efecto, el modelo probabilístico, aun siendo un **modelo binario**, efectúa **equiparación parcial**, lo que permite ordenar los documentos de la respuesta conforme a su probabilidad de relevancia. Ya que no puede ponderar los términos de la colección (es un modelo binario), la equiparación parcial es posible gracias a la ponderación de los términos empleados en la consulta.

A mi juicio, una de las grandes aportaciones del modelo probabilístico a la recuperación de información consiste en el fenómeno denominado **retroalimentación por relevancia**. Aunque con carácter muy general consiste en la utilización de información generada bien en procesos de recuperación anteriores, bien durante el propio proceso de búsqueda, en el modelo probabilístico clásico consiste en mejorar los resultados de la recuperación solicitando al usuario, tras la respuesta inicial del sistema, que analice los documentos recuperados (alrededor de la primera media docena) y juzgue cuáles son relevantes. Con esta información se imponen nuevos valores a las "probabilidades de ser buen y mal descriptor" para cada término de la consulta, obteniéndose una nueva respuesta de documentos ordenados por su probabilidad de relevancia, aunque ahora mejorada, sin duda, gracias a la información suministrada directamente por el usuario.

Actualmente son muchos los sistemas de recuperación de información que emplean alguna variante de la retroalimentación por relevancia para mejorar y refinar los resultados de la búsqueda. Quizá la más conocida consista en la sugerencia al usuario de más resultados precedidos de la siguiente advertencia: “Otros usuarios que adquirieron o preguntaron por ese documento también adquirieron o preguntaron por estos otros”. Es una manera de emplear la información procedente, en este caso, de procesos de recuperación anteriores.

Modelo vectorial

Como hemos observado, se considera que el modelo probabilístico clásico supera al modelo booleano clásico en cuanto que el probabilístico efectúa equiparación parcial mientras que el modelo booleano clásico efectúa equiparación exacta. Sin embargo, ambos siguen presentado una característica negativa: ni el modelo booleano ni el modelo probabilístico tienen en cuenta la frecuencia con la que aparecen los términos de indización dentro de los documentos. Esto es, ambos son modelos binarios de representación documental.

Parece lógico pensar que, si en un documento aparece el término “biblioteca” una vez, y en otro documento aparece ese mismo término veinte veces, consideremos que en el primer documento la importancia de “biblioteca” es menor que ese mismo término en el segundo documento. En consecuencia, surge un tercer modelo de recuperación, el **modelo vectorial**, basado en tres principios[13]:

- La **equiparación parcial**, esto es, la capacidad del sistema para ordenar los resultados de una búsqueda, basado en el grado de **similaridad** entre cada documento de la colección y la consulta.
- La **ponderación de los términos en los documentos**, no limitándose a señalar la presencia o ausencia de los mismos, sino adscribiendo a cada

[13] Sobre el modelo vectorial pueden consultarse Salton, G.; McGill, M. J. Introduction to modern information retrieval. New York: McGraw-Hill, 1983; Meadow, Ch. T.; Boyce, B. R.; Kraft, D. H. Text information retrieval systems. Toronto: Academic Press, 1992; Salton, G.; Wong, A.; Yang, C. S. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613-620.

término en cada documento un número real que refleje su importancia en el documento.

- La **ponderación de los términos en la consulta**, de manera que el usuario puede asignar pesos a los términos de la consulta que reflejen la importancia de los mismos en relación a su necesidad informativa.

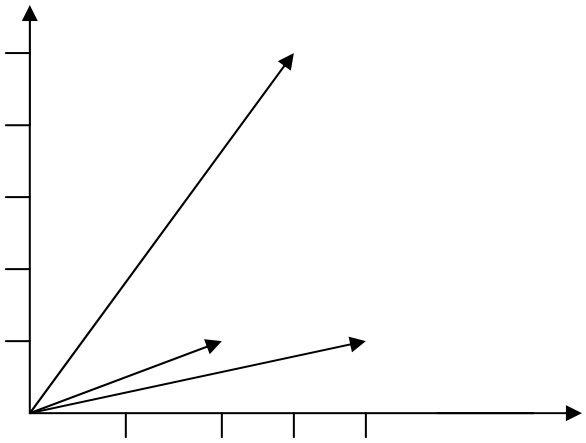
En definitiva, en el modelo vectorial tanto un documento como una consulta se representan mediante conjuntos ordenados de números (no solamente de ceros y unos como en los modelos anteriores; sin embargo, en este modelo también el cero representa la ausencia del término en el documento):

$$\vec{D}_j = (d_{j1}, d_{j2}, ..., d_{jt})$$
$$\vec{Q} = (q_1, q_2, ..., q_t)$$

Donde "t" es el número total de términos considerados en la descripción de la colección. Lógicamente, aunque el usuario no va a introducir nunca "t" términos, se puede representar siempre en función de tales términos (basta imponer ceros en aquéllos no empleados en la consulta).

Gracias a esta representación, tanto los documentos como las consultas pueden tratarse matemáticamente como vectores en un espacio t dimensional, de donde el nombre de **modelo vectorial**. Para que podamos comprender las consecuencias de este hecho, consideraremos únicamente dos dimensiones (esto es, dos únicos términos). Sean, por ejemplo, los documentos D1=(3, 5) y D2=(4,1) y la consulta Q=(2,1). Tratándose de un espacio bidimensional, podemos dibujar tanto los documentos como la consulta en el plano de esta hoja, como "flechas" o vectores que parten del origen de coordenadas, cuyo primer número corresponde al valor del término 1 representado en el eje de abcisas y cuyo segundo número corresponde al valor del término 2 representado en el eje de ordenadas. En nuestro ejemplo obtendríamos el siguiente gráfico, donde D1 es el vector más próximo al eje de ordenadas, D2 es el vector más próximo al eje de abcisas, y donde la consulta Q es el vector entre D1 y D2:

Gráfica 3



Como podemos observar en el gráfico, puede resultar relativamente fácil juzgar cuál de los dos documentos se asemeja más a la consulta. Considerando que el vector de la consulta Q está más próximo a D2, podemos deducir gráficamente que el orden de relevancia de los documentos D1 y D2 en relación a la consulta Q sería en nuestro ejemplo: D2 y posteriormente D1. En resumen, en el modelo vectorial basta fijar un criterio de similaridad para poder ordenar fácilmente por orden de relevancia los documentos de una colección en relación a una consulta.

En cuanto a la manera de ponderar los términos en los documentos de la colección, una de las más utilizadas y de eficacia probada es la denominada **ponderación tf.idf**. Consiste en multiplicar dos factores que reflejan la importancia de los términos:

- El primer factor, **tf (abreviatura de Term Frequency), pretende reflejar la importancia de los términos en los documentos**, concediendo mayor importancia a los términos cuantas más veces aparezcan en los documentos. La versión más sencilla de este factor lo representa numéricamente mediante la frecuencia de aparición de cada término en cada documento de la colección.

- El segundo factor, **idf (abreviatura de Inverse Document Frequency), o inverso de la frecuencia de documentos, pretende reflejar la importancia de los términos en la colección**, primando la precisión y el poder discriminatorio de los mismos. Así, dará mayor importancia a un término cuanto menor sea el número de documentos de la colección en los que aparezca dicho término. Por el contrario, si un término aparece en todos los documentos de la colección, su precisión y poder discriminatorio (capacidad para discernir los documentos relevantes de los irrelevantes ante una consulta) es nulo (tal término aparecerá necesariamente tanto en todos los documentos relevantes como en todos los documentos irrelevantes), de manera que se le otorgará una importancia mínima en esa colección en concreto (puede que en otra colección ese mismo término posea una gran importancia, porque aparece en muy pocos documentos). Suele representarse numéricamente de manera proporcional al logaritmo neperiano del inverso del número de documentos de la colección en los que aparece dicho término.

Como expusimos anteriormente, el modelo vectorial propone evaluar el grado de similaridad entre los documentos de una colección y las consultas mediante algún criterio que muestre la mayor o menor cercanía entre los vectores correspondientes a los documentos y el vector correspondiente a la consulta. Una de las maneras más habituales de cuantificar el nivel de cercanía entre vectores es mediante el coseno del ángulo que forman, pues presenta la propiedad de ser un número mayor cuanto más cercanos estén entre sí ambos vectores (en el límite, el coseno de 0º vale la unidad), mientras que es un número menor cuanto más alejados estén entre sí (en el límite, el coseno de 90º vale cero).

Una vez calculada la similaridad entre cada documento de la colección y la consulta, el sistema es capaz de ordenar todos los documentos de la colección en orden decreciente de su grado de similaridad con la consulta, incorporando de este modo a los resultados aquellos documentos que satisfacen sólo parcialmente los términos de la consulta. Se efectúa, en consecuencia, **equiparación parcial**.

Conclusiones

Actualmente la recuperación de información ha cobrado un gran auge debido al crecimiento espectacular de Internet, tratando de facilitar la tarea de discernimiento de los escasos documentos relevantes que puedan existir en la red frente a los millones de documentos irrelevantes en relación a cada consulta formulada en la red. Dado que esta inmensa "colección" carece por completo de organización, la automatización de los procesos de análisis y recuperación de los billones de documentos que configuran la red se ha convertido en una tarea de importancia capital.

Los programas que rastrean la web en busca de páginas y los programas que efectúan el proceso de análisis y tratamiento de tales páginas con el objeto de poder recuperarlas ante las consultas de los usuarios, además de muchos otros programas con un objetivo semejante en cualquier ámbito (desde las bibliotecas hasta el comercio electrónico), se siguen basando en los tres modelos clásicos de recuperación de información creados entre los años sesenta y ochenta del siglo XX: los modelos booleano, probabilístico y vectorial.

Como hemos podido observar, el fenómeno más destacado actualmente en estos sistemas de recuperación de información consiste en el empleo simultáneo de características y algoritmos propios de cada uno de estos modelos. Así, es muy frecuente que los buscadores de Internet utilicen en un primer paso el modelo booleano, pero empleen posteriormente criterios muy diversos –desde la presencia del término en el título frente al cuerpo del documento hasta la frecuencia con que se actualiza el documento en Internet, por ejemplo- para efectuar la ordenación de los documentos de las respuestas. Ahora bien, sean cuales sean los factores considerados relevantes en la ordenación, ello implica sin duda la utilización en sus algoritmos de los principios del modelo vectorial clásico. De igual modo, cada vez en mayor medida los SRI emplean una u otra variante de la retroalimentación por relevancia para aumentar la precisión de la respuesta, técnica empleada en sus inicios por el modelo probabilístico.

En consecuencia, puede afirmarse que con la popularización de Internet han cobrado nuevo auge los modelos clásicos de recuperación de información, tratando de aunar en un mismo programa de recuperación las ventajas primordiales de cada uno de

ellos. La investigación en este área, muy activa en la actualidad, sigue tratando de mejorar la precisión y exhaustividad de los sistemas, pero tratando ahora de incorporar el usuario real y su punto de vista subjetivo en la evaluación de los sistemas. Sin duda en un futuro –esperemos que no muy lejano- tales avances se incorporarán a los SRI en beneficio de un acceso rápido y eficaz a la información por parte de cualquier habitante de nuestro planeta.

Bibliografía

BAEZA-YATES, R.; RIBEIRO-NETO, B. (1999) Modern information retrieval. New York: ACM, 1999.

CONNOLLY, D. (2000). A little history of the World Wide Web. Disponible en http://www.w3.org/History.html. [Consulta: 16/03/2006].

CROFT, W. B.; HARPER, D. J. (1979). Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 1979, 35(4):285-295.

DIGITAL LIBRARIES INITIATIVE (2006). Disponible en http://dli.grainger.uiuc.edu/national/spanish/index.html. [Consulta: 16/03/2006].

ELLIS, D. (1996) Progress and problems in information retrieval. London: LA, 1996.

GIL, P. How Big is the Internet? (2005). Disponible en http://netforbeginners.about.com/cs/technoglossary/f/FAQ3.htm. [Consulta: 16/03/2006].

GULLI, A.; SIGNORINI, A. (2006). The indexable web is more than 11.5 billion pages. Disponible en http://www.cs.uiowa.edu/~asignori/web-size/. [Consulta: 16/03/2006].

INTERNET Systems Consortium (2006). Internet Domain Survey Host Count. Disponible en http://www.isc.org/index.pl?/ops/ds/. [Consulta: 16/03/2006].

INTERNET World Stats (2006). Internet growth statistics. Disponible en http://www.internetworldstats.com/emarketing.htm. [Consulta: 15/03/2006].

KORFHAGE, R. (1997) Information storage and retrieval. New York: John Wiley & Sons, 1997.

MEADOW, Ch. T.; BOYCE, B. R.; KRAFT, D. H. Text information retrieval systems. Toronto: Academic Press, 1992.

MOYA ANEGÓN, F. de. (1995) Los sistemas integrados de gestión bibliotecaria: estructuras de datos y recuperación de información. Madrid: ANABAD, 1995.

READINGS in information retrieval (1997). (eds.) Sparck Jones, K.; Willett, P. San Francisco: Morgan Kaufmann, 1997.

RIJSBERGEN, C. J. van. (1979) Information retrieval. London: Butterworths, 1979. Disponible en http://www.dcs.gla.ac.uk/Keith/Preface.html [Consulta: 16/03/2006].

ROBERTSON, S. E. (1977) The probability ranking principle in IR. Journal of Documentation, 1977, 33(4):294-304.

SALTON, G.; McGILL, M.J. (1983) Introduction to modern Information Retrieval. New York: Mc.Graw-Hill, 1983.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613-620.

SPARCK JONES, K. (1979) Search term relevance weighting given little relevance information. Journal of Documentation, 1979, 35(1): 30-48.