NeoInstrumenta

Revista del Laboratorio de Ingeniería Documental

Vol. 7
Año 2010

Limitaciones de los modelos tradicionales de Recuperación de Información en la búsqueda Web

Fuensanta Mª Guerrero Carmona
Manuel Marcos Aldón

1. Introducción

El modelo booleano, el espacio vectorial y el probabilístico representan tres problemas clásicos de la Recuperación de Información (RI): consultas estructuradas, peso de los términos iniciales, y realimentación por relevancia, respectivamente.

El modelo booleano es el más utilizado de los modelos de "encaje exacto", en los que la consulta establece unos criterios totalmente precisos de relevancia. Devuelve los documentos que cumplen las restricciones especificadas en la consulta, sin ningún orden o clasificación. Las consultas son expresiones lógicas en las que los operandos son propiedades de los documentos. Los operadores booleanos son la conjunción (AND), la disyunción (OR) y la negación (NOT). Además, la mayoría de los sistemas incluyen operadores de proximidad y expresiones regulares sencillas.

El modelo espacio vectorial, desarrollado por Salton [3], considera el índice y la consulta como vectores en un espacio vectorial multidemensional, dónde cada término del índice es asignado a una dimensión, de forma que la coordenada correspondiente para un documento o consulta indica si ese término está en el documento y con qué peso. La relación entre documentos y consultas se puede deducir comparando sus vectores mediante álgebra lineal.

En los modelos probabilísticos la RI se entiende como un proceso de clasificación. Para cada consulta se establecen dos clases diferenciadas: documentos relevantes y no relevantes. Se calcula la probabilidad de que un documento pertenezca a la clase de los documentos relevantes. El documento se devuelve como resultado si esa probabilidad es mayor que la complementaria, es decir, si P(Relevante|Documento) > P(NoRelevante|Documento). Estas probabilidades se pueden descomponer de forma que se calcula la probabilidad de que cada término de la consulta aparezca en el documento, sabiendo que éste es (o no es) relevante.

A continuación se van a comentar las limitaciones de estos modelos en la búsqueda Web.

2. Limitaciones de los modelos tradicionales de RI

Los principales problemas de la búsqueda y RI en la Web se deben a su tamaño, a la falta de estructura y a su naturaleza dinámica. Además, las consultas suelen ser poco informativas, con una longitud media de aproximadamente 4 palabras. En general, es muy difícil obtener resultados exactos con consultas tan cortas.

La Web se puede considerar un gran repositorio de información distribuida y accesible en permanente crecimiento. Millones de usuarios particulares y organizaciones agregan, eliminan o modifican el contenido de sus sitios Web continuamente, y también puede variar su ubicación. La red contiene documentos de distinta naturaleza y formato, desde páginas HTML hasta archivos de vídeo e imágenes, en formatos estándar y propietario. En un entorno tan heterogéneo, la búsqueda de información se convierte en un gran reto que los sistemas de búsqueda y RI en la Web tratan de superar. Baeza-Yates [1] considera que existen básicamente dos tipos de problemas en este campo: los relacionados con los datos en sí mismos y los relativos a los usuarios.

Los sistemas tradicionales de almacenamiento y RI suelen ser utilizados por profesionales que conocen la estructura de la colección y dominan el uso de los operadores booleanos. Por lo tanto, el número de documentos recuperados depende de la utilización correcta de dichos operadores y de la selección adecuada de los descriptores de búsqueda. Cuando los modelos booleanos se utilizan para acceder a bases documentales tradicionales de contenido muy especializado, es posible recuperar **toda** la información pertinente, contando siempre con la experiencia y habilidad del profesional que realiza la consulta.

Pero en los sistemas de búsqueda y RI actuales, como los motores de búsqueda Web, las premisas anteriores no se cumplen: la mayoría de los usuarios no dominan el lenguaje de los operadores booleanos, y la "colección" de documentos no tiene una estructura determinada, no está expresada en el mismo idioma y a veces no tiene la calidad adecuada. Además, para una consulta media de 3 ó 4 palabras se pueden recuperar miles de documentos. Pero, normalmente, un usuario sólo accede a los diez documentos que se presentan en primer lugar, por lo que la clasificación de los resultados según la relevancia para el tema que se está buscando es de vital importancia. Puesto que los modelos booleanos no implican una ordenación de los resultados, no son operativos en los sistemas de búsqueda y RI en la Web. No obstante, se utilizan como base en la mayoría de los motores de búsqueda, que admiten tanto la búsqueda booleana como la búsqueda en lenguaje natural.

Los modelos de espacio vectorial y probabilísticos utilizan un tipo de estructuras que permiten acceder eficientemente a los datos procesados: los índices inversos, en los que se almacena cada término junto con el número de ocurrencias (peso) y la dirección donde está ubicado, además de otra información complementaria (si está en la cabecera o en el título, si está en negrita, etc.). Los datos obtenidos de los índices sirven para clasificar y ordenar los resultados según su relevancia, que se puede llevar a cabo:

- Utilizando términos: ya sea por su frecuencia (cantidad, cercanía, ubicación), o por la similitud de documentos.
- Utilizando hipervínculos: ranking de popularidad, hubs y autoridades.

En el modelo de espacio vectorial se determina la similitud que existe entre una consulta y un documento calculando la distancia entre los vectores que los representan: mayor número de términos coincidentes, dando preferencia a aquellos documentos en los que los términos de la consulta aparezcan más veces. Pero en la Web, medir los pesos de todos los términos que aparecen en todos los documentos es una tarea imposible de llevar a cabo. Indexar la Web completa es inviable, por su tamaño, por la volatilidad de su contenido, porque existen datos redundantes y obsoletos, y por su heterogeneidad (archivos multimedia, idiomas diferentes, etc.). Además, las páginas Web dinámicas (contenido XML, acceso a través de formularios o páginas protegidas por contraseña, etc.) no permiten la indexación de toda la Web, por lo que la exhaustividad y cobertura no es completa. La precisión de los resultados obtenidos tampoco es la deseable, ya que el modelo de espacio vectorial presupone la independencia entre los términos de indexación. En el entorno Web, donde la mayoría de las consultas se realizan en lenguaje natural, esta característica provoca la obtención de falsos negativos (no se recuperan documentos que son relevantes; por ejemplo, barato, económico, asequible, etc. son sinónimos, pero no coinciden) y falsos positivos (se recuperan documentos que no son relevantes porque coincide parte del término).

Por su parte, los modelos probabilísticos utilizan los índices inversos de forma adaptativa, calculando la probabilidad de relevancia de un documento con respecto a una consulta y ordenando los resultados según dicha probabilidad. Además, se valora la interacción con el usuario y lo que éste considera relevante para su necesidad de información. En este sentido, se mejoran los

resultados mediante la retroalimentación por relevancia, que consiste en la utilización de información generada bien en procesos de recuperación anteriores, bien durante el proceso de búsqueda actual. Se tienen en cuenta no solo los contenidos, sino los datos almacenados sobre cada usuario, los sitios visitados, el tiempo de permanencia, etc. Para ello son fundamentales los datos que proporciona la estructura de enlaces de la Web. Los hipervínculos son una fuente valiosa de información sobre los usuarios, sus preferencias, sus necesidades e intereses. También los hubs y autoridades [2] aportan conocimiento sobre la red. La autoridad calcula el valor que tiene el contenido de una página, señalando fuentes destacadas y con calidad probada. Los hubs evalúan la importancia de los enlaces de una página hacia autoridades o sitios destacados y significativos sobre un tema particular. Toda esta información que aportan los enlaces es tenida en cuenta a la hora de ordenar los resultados que se presentan al usuario y para asignar mayor probabilidad a los documentos que, posiblemente, satisfagan sus necesidades de búsqueda.

Pero esta relevancia o importancia de las páginas se puede manipular mediante una técnica llamada *spamdexing*, que afecta tanto al contenido como a los enlaces y, en consecuencia, a los resultados obtenidos con modelos de espacio vectorial y probabilísticos. En relación al contenido, algunos métodos utilizados son: incluir texto oculto o invisible, colocar palabras clave dentro de una página para elevar el conteo de estos términos, relleno de de metaetiquetas con palabras clave que no tienen relación con el contenido real de la página, etc. El spamdexing de enlaces afecta a algoritmos de posicionamiento como PageRank y HITS, que asignan una posición más alta a los sitios que son enlazados por muchos otros, relacionado con los conceptos de autoridad y hub mencionados antes. Algunas de las técnicas utilizadas son: granja de enlaces (comunidades de páginas que se referencia mutuamente para conseguir mejor posición), enlaces ocultos, *Sybil attack* (creación de sitios falsos que se enlazan entre sí), etc.

Todos estos problemas que se presentan en la RI de la Web, sumados a los mencionados antes (tamaño, volatilidad, falta de estructura, datos redundantes y obsoletos y heterogeneidad), imponen ciertas restricciones a las herramientas de búsqueda en cuanto a la cobertura y acceso a los documentos, exigiendo cada vez mayores recursos computacionales, como espacio de almacenamiento, ancho de banda o velocidad de procesamiento, además de la necesidad de utilizar diferentes estrategias para mejorar la calidad de las respuestas. Para lograr este objetivo, actualmente se utilizan variaciones de los modelos tradicionales de RI, como los basados en lógica difusa, indexación semántica latente, redes de inferencia, etc.

Referencias

[1] BAEZA-YATES, R., RIBEIRO-NETO, B. "Modern Information Retrieval - the concepts and technology behind search". Segunda Edición. Pearson Education Ltd., Harlow, England, 2010.
ISBN: 9780321416919

[2] KLEINBERG, J. M. “Hubs, authorities, and communities”. En *ACM Computing Surveys (CSUR)* [en línea]. Vol. 31, nº 4es, (1999). Disponible en: http://cs.brown.edu/memex/ACM_HypertextTestbed/papers/10.html

[3] SALTON, G., MCGILL, M. J. "Introduction to modern information retrieval". McGraw-Hill computer science series, 1983.
ISBN: 9780070544840