Uso real de los modelos matemáticos en los motores de recuperación de la información

Jordi Ardanuy

Departament de Biblioteconomia i Documentació
jordi_ardanuy@ub.edu
Tel. 93 403 47 53 - Fax. 93 403 5772
Universitat de Barcelona
Melcior de Palau, 140. 08014 Barcelona

Resumen

La recuperación de la información exige la utilización de modelos matemáticos que comparen las peticiones del usuario con el contenido de los documentos.

Este trabajo persigue conocer cual ha sido la implementación real, y no meramente experimental o teórica, de los modelos matemáticos utilizados en los diversos motores de recuperación de la información de propósito general que pueden obtenerse en la actualidad.

Palabras Clave: Recuperación de la información. Sistemas de recuperación de la información. Motores de búsqueda. Modelos de recuperación de la información.

1 Introducción

La recuperación de la información (RI) trata sobre la representación, almacenamiento, organización y recuperación de documentos. Tal proceso debe dotar al usuario de un acceso funcionalmente sencillo a la información en la que pudiera estar interesado. Aunque la disciplina es mucho más amplia, nosotros nos ceñiremos aquí a los sistemas automáticos de recuperación de documentos de texto no estructurados.

La tarea de caracterizar las necesidades informativas del usuario es un problema complejo, generalmente abordado mediante la traducción a una expresión más o menos sintética, llamada ecuación, que debe ser procesada por un motor de búsqueda.

La ecuación de consulta suele estar constituida por un conjunto de palabras clave o términos de indexación, que deben representar las necesidades informativas del usuario. Por su parte, los documentos también son representados por colecciones de términos. Comparando la ecuación suministrada por el usuario con los esquemas que representan los documentos se

pueden elegir aquellos que sean más pertinentes a su demanda. El proceso de búsqueda puede refinarse mediante un proceso de retroalimentación.

La representación y comparación entre los términos de los documentos y de las consultas se realiza siguiendo un modelo matemático. Existen diversas propuestas teóricas suficientemente documentadas. Sin embargo, estas fuentes no recogen cuántos y cuáles han sido los modelos que se han implementado realmente, ni enumeran de forma extensiva los motores de RI aplicables a diversos propósitos (bases de datos documentales, páginas web,…) que pueden obtenerse bajo algún tipo de licencia en la actualidad. El objetivo de nuestro trabajo consiste en responder a dichas preguntas, mostrando además en qué medida el uso de unos u otros modelos ha variado con el tiempo.

Para ello se ha recurrido a una exploración exhaustiva de la bibliografía especializada, la consulta a expertos en sistemas de recuperación de la información, los recursos en línea que se ofrecen al respecto desde diversas instituciones científicas, así como la documentación que generan las organizaciones o individuos que han desarrollado los ingenios de RI. Se trataba de conseguir, en primer lugar, establecer un listado lo más completo posible de ingenios que estuvieran operativos y disponibles para, posteriormente analizar qué tipo de modelo matemático utilizan. La documentación técnica necesaria para nuestro estudio estaba en diversos casos previamente publicada, aunque no siempre de manera sistemática. En otros ha sido necesario establecer contacto con quienes han intervenido en el desarrollo del proyecto para que subsanaran las lagunas informativas.

El esquema del texto es como sigue. Primero se enumeran los diferentes modelos matemáticos propuestos para uso en motores RI y su agrupación en clases. Posteriormente sintetizamos los resultados de nuestra investigación. Finalmente ofrecemos las conclusiones.

2 Modelos conceptuales en RI

Una taxonomía ya clásica agrupa los diferentes esquemas matemáticos en tres bloques: los basados en la teoría de conjuntos, los algebraicos y los probabilísticos.

Algunos de ellos no han pasado de ser un aparato teórico, mientras que otros se han quedado en los laboratorios. Finalmente una parte ha sido implementada, sin que se pueda hablar de un claro vencedor.

2.1 Modelos de teoría de conjuntos

El primer modelo adoptado por los sistemas automáticos de RI se basa en el álgebra de Boole. En él se considera que los términos indexados están simplemente presentes o no. Este

esquema de decisión binario no contempla ninguna graduación, por lo que se han propuesto otras soluciones basadas en la lógica difusa [1] mejor adaptadas a las relaciones entre conceptos poco definidas, así como combinaciones con modelos algebraicos [2] para presentar al usuario una colección de referencias jerarquizada (ranking).

2.2 Modelos algebraicos

El modelo más antiguo y arraigado de este grupo asimila el conjunto de términos que caracterizan la colección con un espacio vectorial, en el cual la dimensión viene dada por el número de términos diferentes que representan a los documentos, frecuentemente muy grande. Cada vector describe un documento, y el valor de las componentes representa la presencia ponderada de cada término en el documento. Existen diversas variantes [3]. La consulta también se estructura como un vector de dicho espació. Para comprobar el grado de similitud entre los vectores que representan la colección y el de la consulta se utiliza algún tipo de correlación que permite un orden jerárquico de relevancia.

El modelo de indexación semántica latente (LSI, acrónimo en inglés) adopta una perspectiva diferente, al hacer aflorar los conceptos que contiene el documento, más que preocuparse por los términos que aparecen. El resultado es un vector en un espacio de dimensión reducida. La principal técnica utilizada se conoce como descomposición en valores singulares [4].

2.3 Modelos probabilísticos

En este caso los algoritmos se basan en estimar la probabilidad de que un documento relevante a un usuario pueda recuperarse mediante una ecuación de búsqueda concreta.

El modelo más sencillo [5], es el llamado de independencia binaria (BIR, acrónimo en inglés) en el cual se admite la independencia de cada término. El cálculo de la similitud entre la consulta y el documento se realiza mediante la razón de proporcionalidad entre la probabilidad de que el documento sea relevante y que no lo sea (función odd), calculada usando Bayes, a partir de la presencia de los términos.

Esquemas más avanzados se basan en asociar variables aleatorias a términos, documentos y consultas de los usuarios que constituyen nodos de grafos dirigidos. Es el caso de las redes de inferencias bayesianas [6] y el modelo de red de creencias [7].

Un planteamiento muy diferente lo ofrece el modelo de regresión multivariante logístico por etapas (SLR, acrónimo en inglés) [8], que opera a partir de datos adquiridos en un proceso previo de entrenamiento con algunos documentos de una colección.

3. Sistemas de RI en la actualidad

El presente estudio abarca los sistemas de RI integrables en bases de datos documentales, y por tanto no circunscritos a un producto concreto, de los cuales es posible obtener licencias, de pago o gratuitas, en la actualidad. Estos son los diecisiete sistemas analizados: Smart, Personal Librarian (SIRE), OKAPI, Xapian (Omsee), Smartlogic Discovery, Glimpse, Telcordia LSI, Inquery, Managing Gigabytes (MG), Isearch, IB, Cheshire II, Prise, LSI ++, Verity Ultraseek, GTP y Amberfish.

Si los clasificamos en función del grupo al que pertenece el modelo matemático que utilizan, existe un claro predominio de los algebraicos, que son aproximadamente dos terceras partes (Tabla 1). No llegan a un 30 % los de tipo probabilístico. Un caso excepcional es el GLIMPSE, que utiliza una búsqueda meramente secuencial.

Aunque casi todos los sistemas admiten ecuaciones con operadores de Boole, constituye una herramienta accesoria y no la estrategia central utilizada por el motor de búsquedas, lo que contrasta, por ejemplo, con el modelo seguido por motores como los usados por Google o Altavista, que se centran exclusivamente en la lógica de Boole.

Familia	Nº de sistemas	Porcentaje
Algebraicos	11	65 %
Probabilísticos	5	29 %
Secuenciales	1	6 %

Tabla 1

Si nuestra atención ahora recae en los modelos concretos (Tabla 2), el de espacio vectorial es el utilizado en casi la mitad de los sistemas, frente al probabilístico independiente (BIR) y el de indexación semántica latente (LSI), que aparecen en la quinta parte en ambos casos. Se nota en los años 90 un desplazamiento hacia modelos más sofisticados que los relativamente elementales vectorial o BIR.

Ningún motor de búsqueda utiliza un modelo basado en lógica difusa ni otras posibilidades propuestas en el ámbito de la teoría de conjuntos. Tampoco existe ningún sistema que utilice la propuesta teórica más moderna (1996) del modelo de red de creencias.

Modelo	Nº de sistemas	Porcentaje
Espacio vectorial	8	47 %
BIR	3	18 %
LSI	3	18 %
Red de inferencia	1	6 %
Regresión logística	1	6 %
Secuenciales	1	6 %

Tabla 2

Las tabla 3 relaciona cada motor con el modelo matemático que incorpora, detallándose también la fecha de aparición de su primera versión, si continúa actualizándose, si tiene propósitos comerciales, si existe versión gratuita y una URL de referencia (diciembre de 2003).

Sistema y URL de referencia	Modelo de recuperación	Año 1ª versión	Proyecto activo	Aplicación comercial	Versión gratuita
Smart (versión moderna sobre UNIX) <ftp.cs.cornell.edu/pub/smart>	Vectorial	1981(*)	No	No	Sí
Personal Librarian (SIRE) <http://www.pls.com/downinst.htm>	Vectorial	1983-1986	No	Sí	No
OKAPI <http://www.soi.city.ac.uk/~andym/OKAPI-PACK/registration_details.php>	BIR	1984	No	No	No
Xapian (Omsee) <http://xapian.sourceforge.net/download.php>	BIR	1984	Sí	No	Sí
Smartlogic Discovery <http://www.aprsmartlogik.com/ products/discover>	BIR	1984	Sí	Sí	No
Glimpse <http://glimpse.cs.arizona.edu>	Búsqueda secuencial	1993	Sí	Sí	Sí
Telcordia LSI <http://lsi.argreenhouse.com/lsi/request.html>	LSI	1994	Sí	Sí	No
Inquery <http://ciir.cs.umass.edu>	Red d'inferencia	1994	Sí	Si	No
Managing Gigabytes (MG) <http://www.cs.mu.oz.au/mg/mg-1.2.1.tar.gz>	Vectorial	1994	No	No	Sí
Isearch <http://www.freesoft.org/software/Isearch>	Vectorial	1994	No	No	Sí
IB <http://www.bsn.com/main.html>	Vectorial	1995	No	Sí	No
Cheshire II <ftp://cheshire.berkeley.edu/pub/cheshire>	Regresión logística	1995	Sí	No	Sí
Prise <http://www-nlpir.nist.gov/projects/zprise/index.html>	Vectorial	1995	No	No	Sí
LSI ++ <http://www.cs.utk.edu/~lsi/request2.html>	LSI	1996	No	No	Sí
Verity Ultraseek <http://downloadcenter.verity.com/dlc/index.jsp>	Vectorial	1997	Sí	Sí	Sí (30 días)
GTP <http://www.cs.utk.edu/~lsi>	LSI	1998	Sí	No	Sí
Amberfish <http://www.etymon.com/amberfish>.	Vectorial	2002	Sí	Sí	Sí

Tabla 3

A partir de los datos cronológicos de la tabla 3 podemos estudiar la evolución de implementación de modelos (Figura 1), observando cómo el esquema vectorial ocupa la primera mitad de los 80, recuperándose el interés por ellos hacia la mitad de los 90 coincidiendo con la expansión de la www. Sin embargo, hay que considerar que el SMART, y en menor medida el SIRE, estaban vivos como proyectos de investigación con anterioridad.
El momento de máximo desarrollo del BIR se sitúa a principios de los 80, mientras que el LSI adquiere protagonismo a partir de la mitad de los noventa.

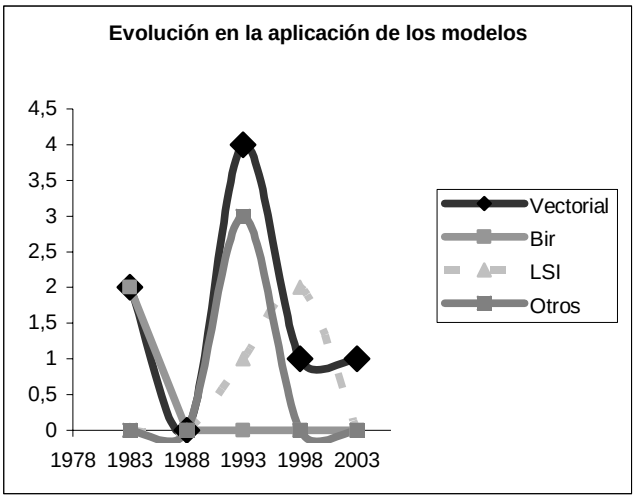

Figura 1

Si analizamos finalmente el desarrollo cronológico de la producción de sistemas de RI (figura 2), se pone de manifiesto que, si bien ha existido una tendencia continua a la producción, existen dos etapas especialmente fructíferas, una hacia la mitad de los años 80, justo coincidiendo con la implantación del PC y la segunda, en el momento de la eclosión informativa en Internet.

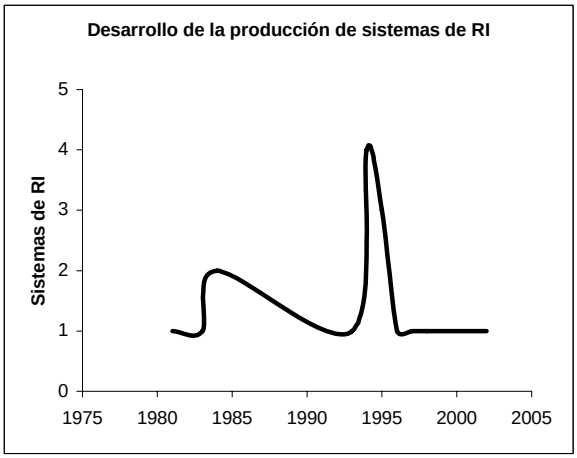

Figura 2

Pasado el fuerte impulso de productos desarrollados a mitad de los 90, la producción de motores ha decaído. Las razones hay que buscarlas, por una parte, en que ya existe un abanico de motores suficientemente abundante para satisfacer las demandas esenciales. Por otro lado, no existen fundamentos teóricos de suficiente peso para justificar la idea que propuestas más sofisticadas como el modelo de red de creencias fueren a mejorar significativamente el rendimiento. Las confrontaciones experimentales entre los motores actuales tampoco han animado a pensarlo. Los bancos de pruebas comparativas a los que se les ha sometido frente a las mismas colecciones de documentos, como en el caso de las TREC (http://trec.nist.gov/pubs.html), no han arrojado vencedores suficientemente nítidos, ni siquiera contra los modelos de más sencilla implementación como los de tipo vectorial, lo que les sigue proporcionado vigencia.

Esta timidez en la producción de los últimos años nos llevó a estudiar las nuevas tendencias en RI. Más que producción de nuevos motores con nuevos modelos matemáticos, las líneas de investigación y desarrollo añaden esquemas alternativos con fuerte base lingüística, como en el caso del sistema PIRCS (http://ir.cs.qc.edu/pircs.html). El interés se ha focalizado hacia nuevos desafíos como la recuperación multilingüe, la de imágenes o la de textos estructurados.

Finalmente se trabaja con especial intensidad en la mejora de la comunicación humano-máquina (IHM) de manera que pueda favorecer el proceso de retroalimentación y permitir superar las limitaciones intrínsecas a todo modelo lingüístico-matemático de RI.

4. Conclusiones

Gracias al estudio sistemático realizado disponemos de un listado de los 17 motores de RI de propósito general, de los que es posible obtener actualmente algún tipo de licencia. El análisis de los modelos matemáticos que aplican nos ha permitido determinar qué propuestas teóricas se han implementado y cuáles no. Se ha comprobado que el modelo más extendido es el de tipo vectorial, especialmente por su sencillez. Fue el primero en introducirse en los años 80 al comprobar las limitaciones del álgebra de Boole.

En los años 90 han aparecido motores con nuevos esquemas como la indexación semántica latente (LSI) o las redes de inferencia que son más sofisticados, pero que pese a ello no han arrojado resultados sensiblemente mejores, lo que ha permitido sobrevivir cómodamente al modelo vectorial.

Sabemos ahora que existen dos periodos de gran productividad de motores de RI, que se corresponden con la extensión de los ordenadores personales y las bases de datos electrónicas

en un caso, y la aparición de la www en otro. Nuestro análisis cuantitativo permite valorar la importancia relativa de tal productividad.

En la actualidad, los modelos matemáticos no son los protagonistas en los esfuerzos de mejora del rendimiento en la RI, debido a la falta de expectativas sensiblemente positivas frente a nuevos modelos.

Referencias

1 Y. Ogawa, T. Morita, and K. Kobayashi (1991). «A fuzzy document retrieval system using the keyword connection matrix and its learning method». Fuzzy Sets and Systems, vol. 38 (1991), pp. 17-41.

2 G. Salton, E. Fox, H. Wu (1983). «Extended Boolean information retrieval». *Communications of the ACM*, vol. 26, n 1 (November 1983), p. 1022-1036.

3 G. Salton, C. Buckley (1988). «Term-weighting approaches in automatic text retrieval». *Information Processing and Management*, vol. 24, n.5 (1988) p. 513-523.

4 S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer (1990). «Indexing by latent semantic anilysis», *Journal of the American Society for Information Science*, vol 41, n. 6 (September 1990), p. 391-407.

5 N. Fuhr (1992). «Probabilistic models of information retrieval». *Computer Journal*, vol. 35, n. 3 (1992) p. 244-255.

6 H. Turtle (1991). *Inference Networks for Document Retrieval.* Ph. D. dissertation. [Amherst]: University of Massachusetts. Disponible en línea <http://citeseer.nj.nec. com/turtle91inference. html> [Consulta: 1 de mayo de 2003].

7 B. A. Ribeiro-Neto, R. Muntz (1996). «A belief network model for IR». **En:** *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieva*l. New York: ACM Press, p. 253-260

8 W. S. Cooper, F.C. Gey, D.P. Dabhey. (1992). «Probabilistic Retrieval Based on Staged Logistic Regression». **En:** *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York: ACM Press, p. 198-210.