Coincidencia y equiparación en los modelos de recuperación de información

Celia CHAIN NAVARRO

Facultad de Comunicación y Documentación
Universidad de Murcia

RESUMEN
En este trabajo se analizan las formas en las que los modelos de recuperación de información más conocidos consiguen hacer coincidir pregunta y respuesta según el paradigma en el que se basan. Para ello se estudia si realmente se persigue que los términos de la pregunta y los de los documentos recuperados coincidan, es decir, sean exactamente iguales, o si por el contrario, se pretende buscar una aproximación semántica entre ambas partes. Si buscamos la coincidencia exacta, el concepto de equiparación es total, mientras que otras formas de aproximarse a ella parcialmente pueden estar basadas en la similitud, o similaridad, pregunta-respuesta o documento-documento; en la probabilidad de relevancia de los documentos frente a las preguntas, en la adyacencia o en la proximidad en el espacio documental, en las co-palabras, co-citación, o co-sitación, entre otras. Se concluye, que mientras el modelo booleano busca la coincidencia exacta, el resto muestran una tendencia a la equiparación pregunta-respuesta mediante fórmulas diferentes.

Palabras clave: Coincidencia, equiparación, modelos de recuperación de información, modelo booleano, modelo vectorial, modelo probabilístico, enfoque cognitivo, modelo fuzzy, relevancia, similitud.

Coincidence and coherency in Models of Information Retrieval

ABSTRACT
This paper is aimed at analysing the procedures by which the most renowned information retrieval models allow to lit question and answer according to underlying paradigms. It becomes so far crucial to establish a priori whether is searched for the adequacy between the terms for question and retrieved documents, or conversely, a semantic approach between both parts. In looking for exact match, the notion of fitness appears rather complete. However, there remain a pleyade of aditional possibilities, such as those based in similarity, question-answer or document-document, probability of relevance of documents against question, adjacency or nearbiness within the documental space, in co-wording, co-citing, co-sitation, among others.It is concluded that, while the Boolean model search for exact matching, the remaining approaches show a tendency to cohere question and answer through different formulae.

Key words: Matching, information retrieval models, boolean model, vector model, probabilistic model, cognitive model, fuzzy model, relevance, similarity.

1. INTRODUCCIÓN

La recuperación de información (RI), como materia en desarrollo dentro de otra discplina más amplia, la Información y Documentación, ha sido uno de los temas que mayor atención ha tenido en los últimos tiempos, especialmente desde

que Salton publicara su libro titulado *Introduction to Modern Information Retrieval* en el año 1983. Sin embargo, las características intrínsecas de la recuperación de información hacen de ella una línea de trabajo interdisciplinar, que fácilmente alcanza la transdisciplinariedad una vez desarrollada; lo que supone que sea necesario realizar un trabajo conjunto con profesionales de disciplinas limítrofes, o al menos complementario, para alcanzar su consolidación. Esta necesidad la hace dependiente, lo que significa que sólo el avance en colaboración, o la especialización transdisciplinar, podrán hacer posible que la recuperación de información sea cada vez más una realidad que no se quede en la metáfora magistral y mágica del sistema informático y en los modelos matemáticos que con frecuencia la alejan del estudio de muchos profesionales de la información, en la creencia sublime de que serán los informáticos y los matemáticos los que solucionen esos problemas.

Paradójicamente, el sencillo planteamiento del **objetivo básico** de la recuperación de información: **que toda la información recuperada sea la información relevante necesaria (o que toda la información relevante sea la recuperada) para satisfacer una consulta**, se complica en exceso cuando intentamos conseguir llevar a la práctica semejante objetivo.

En cualquier conjunto sistemático o no de datos e información (sea una base de datos, el texto de una monografía, o una sede web) el simple hecho de formular una pregunta y buscar en aquéllos la información pertinente, se puede convertir en una tarea francamente difícil. A primera vista, varios son los factores que pueden hacer que la recuperación de la información pertinente sea complicada:

* Que el propio sistema de recuperación no sea fácil de utilizar, ni conocido, ni accesible.
* La pregunta puede estar mal formulada, y no «corresponder» con el lenguaje utilizado en la indización previa de esos documentos.
* Cuando no buscamos una fecha o nombre concreto, sino que queremos recuperar todos los documentos relacionados con la pregunta, no sólo un texto, o párrafo; sobre todo porque muchos de ellos pueden ser complementarios.
* Cuando queremos varios documentos pertinentes, pero no queremos que aparezcan documentos que no responden a la cuestión planteada.
* Cuando los propios documentos, párrafos o datos recuperados con las características arriba citadas no están en un formato adecuado que necesita el usuario.

Esta situación inicial se puede ver agravada por otros factores, como el uso de diferentes idiomas en las operaciones de indización y búsqueda, o la posibilidad de que no coincidan los términos utilizados en ambas operaciones (búsqueda y recuperación).

Por ello, aquí vamos a analizar las formas en las que los documentos se pueden hacer coincidir o equiparase con las preguntas correspondientes según los paradigmas en los que se basan los modelos de RI.

2. PREGUNTAS Y DOCUMENTOS EN LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

Sin ánimo de establecer ninguna teoría sobre la información documental, ni sobre lo que es o no documentación, sí que consideramos oportuno explicar, ampliando o delimitando, el concepto de documento en un sistema de recuperación de la información, para poder plantear posteriormente la posibilidad de identificación, o no, de la pregunta y el documento.

Un documento, tal y como se concebía en la teoría otletiana, es tanto un artículo, como un capítulo de un libro, el mismo libro, un cuadro, un mensaje, una carta, imágenes, gráficos, sonido, ficheros de datos, programas de ordenador, etc. Pero también dentro de un libro, podemos generar varios documentos con cada uno de sus capítulos. Es decir, que la unidad viene establecida por la necesidad de estructurar la información, no por la unidad física que contiene esa información.

Queda ahora la duda sobre si una pregunta de las que realizan los usuarios se puede considerar entonces un documento. La respuesta, que a primera vista puede no parecer que tenga mucha importancia, a posteriori comprobamos que sí lo tiene ya que, dependiendo de la respuesta, terminaremos modelando o estructurando el sistema de recuperación de la información de una forma u otra. Por ejemplo, en los sucesivos encuentros TREC, las preguntas se formulaban en forma de temas, algunos de ellos, por la explicación que precisaban, eran a veces más amplios que algunos de los documentos recuperados. Posteriormente, las preguntas formuladas se incorporaban a la base de datos de documentos por si podían ser de ayuda a posteriores consultas. Aquí la pregunta se consideraba un documento.

Así, una pregunta puede ser considerada, o no, un documento que un sistema de información pueda manejar. En cualquiera de los casos, una pregunta es un tipo de «documento diferente» de los que el usuario suele intentar recuperar. La mayoría de las preguntas son relativamente breves, su lenguaje suele estar restringido y, frecuentemente, no cumplen las reglas normales de sintaxis. Bollman[1], por ejemplo, dice que las preguntas son mucho más cortas que los documentos, que no están formuladas lingüísticamente como los documentos, y que tienen entidad propia.

Si el objetivo es definir una coincidencia entre una pregunta dada y aquellos documentos que al usuario le gustaría recuperar para responder a esa cuestión, el proceso de coincidencia se complica por el hecho de que la pregunta y los documentos pueden tener formas diferentes.

A pesar de lo dicho anteriormente, tampoco debemos olvidar que los documentos y las preguntas, a menudo, sufren procesos paralelos dentro del sistema de recuperación de la información.

---

[1] BOLLMAN-SDORA, P. & RAGHAVAN, V.V. On the delusiveness of adopting a common space for modelling IR objects: Are queries documents?. *JASIS*, 1993, 44, 10, p. 579-587.

3. CRITERIOS DE COINCIDENCIA Y/O EQUIPARACIÓN

Si la pregunta es considerada como un documento, entonces el/los documentos que la respondan serán coincidentes con ella. Si la pregunta *Q* se considera documento, entonces el proceso de recuperación llegará a ser una selección de documento/s que la respondan. Por el propio concepto, no es lo mismo hacer coincidir que equiparar, la última puede llevar implícito una serie de transformaciones de los documentos o de la estructura, que pueden ser (o no) coincidentes con la pregunta. Así, no es lo mismo comparar dos paellas terminadas que establecer similitudes entre una ya hecha y otra en la que puedo cambiar la calidad o el orden de los ingredientes. Esa diferencia es la que existe entre modelar un sistema de RI de una u otra forma.

El usuario siempre espera una respuesta exacta y perfecta a su pregunta. Así que, ésto sólo es posible cuando la pregunta que se hace puede ser identificada definitivamente como que se puede responder, o no.

La coincidencia exacta más obvia suele ser cuando son consultas numéricas. Los términos en cada base de datos son organizados en campos definidos, y es posible determinar exactamente dónde el valor de un término dado coincide con el valor especificado en la pregunta. Todos los datos cuyos valores coincidan con los valores especificados en la pregunta, serán recuperados, y todos los demás rechazados.

Otra posibilidad es que haya coincidencia, pero sobre un intervalo previamente definido en la pregunta, no los números o términos exactos. Entonces se debe ofrecer un rango de valores aceptables para cada término dado, por ejemplo, la coincidencia con la pregunta requiere encontrar dentro de un intervalo determinado años, por ejemplo, «documentos sobre el SIDA publicados entre los años 1995 y 2002». Entonces, la coincidencia por rango es posible en los términos que tienen un orden natural (numérico o alfabético), cuyo significado especifique que un término es mayor que un valor mínimo y menor que un valor máximo. Esta coincidencia también es posible definiendo un conjunto de valores de términos aceptables, y, así, el sistema puede determinar dónde un valor dado está dentro o no del conjunto.

La coincidencia de un término o de una frase frente a otra, es una parte fundamental en el proceso de RI. Sin embargo, el hecho de que una cadena de caracteres de un documento coincidan con un término o una pregunta no significa que automáticamente ese documento recuperado responda a la pregunta formulada. Muchos son los argumentos que sostienen esta afirmación. Entre ellos, aquí señalamos cuatro:

- En primer lugar, porque posiblemente la pregunta contenga más de un término, con lo cual, debe considerarse que debe existir coincidencia entre cada uno de los términos de la pregunta.
- En segundo lugar, el hecho de que un documento contenga un término dado no significa que el documento profundice en lo relativo al término. De hecho, puede ser que esté mencionado «de paso», o también puede ser que el documento contenga un término coincidente, pero que el contexto no sea el apropiado para la pregunta (por vocabulario, idioma, formato, etc.).

– En un caso extremo, incluso podemos encontrar documentos en los que aparezca el término buscado, pero precisamente negando que se hable del contenido, con una frase como, por ejemplo, «este documento no trata sobre», «los efectos de (…) serán tratados en sucesivos artículos».
– Finalmente, un documento, aunque contenga términos coincidentes, puede que no sea útil ni pertinente a la consulta realizada, simplemente porque ha quedado obsoleto o porque el usuario ya lo conoce.

Un sistema es capaz de procesar documentos a texto completo, pero puede ser que la pregunta que se le formule tenga características mixtas. Por ejemplo, un usuario solicita información sobre financiación de proyectos de investigación, que ofrezcan al menos 100.000 euros y que estén dentro del año fiscal 2004. Entonces, el sistema de recuperación de la información, deberá orientarse hacia la coincidencia aproximada basada en el texto, que sólo será pertinente si incluye los componentes marcados en el rango.

La mayoría de los textos e imágenes de las bases de datos o de Internet no se pueden adaptar a este sistema. Entonces, se ha desarrollado un concepto de coincidencia parcial, que Moya denomina **equiparación**[2]. Ésto requiere una forma de medir cómo y cuánto un documento dado «coincide» (se aproxima) con la pregunta. En palabras suyas «La equiparación establece qué es común entre las representaciones de los documentos y las de las búsquedas»[3].

Así cuando equiparamos los contenidos de un documento y los de una pregunta no se busca que los términos de la pregunta y los de los documentos recuperados coincidan, es decir, sean iguales, sino se trata de buscar métodos válidos para definir cuándo un documento *D* es identificado como relativo a una pregunta *Q*. El problema es determinar dónde la relación es lo suficientemente fuerte como para garantizar la recuperación del documento. En este trabajo nos centramos en el grado de aproximación temática del documento, referido a cuándo y cómo el tema del documento es equiparable con el tema de la pregunta. Tal y como lo representa Rijsbergen:

$$Q\_D\_Q \; ? \; 0$$

y teniendo en cuenta que la relación expresada con esa flecha puede quedar definida de muchas formas, y, según cada uno de los modelos existentes, será coincidencia, similitud, posibilidad, probabilidad o grado de equivalencia, etc. Para expresarlo de forma general utilizamos el sustantivo «equiparación» refieriéndonos a la relación expresada en cualquiera de los modelos de RI para definir la relación existente entre la pregunta *Q* y el documento *D* que puede servir de respuesta.

Según Belkin y Croft[4], no es lo mismo la coincidencia (o equiparación) total (el « total matching», en el que sí y sólo si coinciden los términos de la pregunta y del

---

[2] MOYA ANEGÓN, F. Sistemas avanzados de recuperación de la información. En: LÓPEZ YEPES, J. (ed.). *Manual de Ciencias de la Documentación*. Madrid: Pirámide, 2002, p. 573 y ss.

[3] Idem, p. 574.

[4] BELKIN, N. & CROFT, B. Retrieval Techniques. En *Annual Review of Information Science and Technology*, 1987, 22, p. 109-145.

documento hay recuperación pertinente), que el concepto de equiparación parcial («parcial matching»). En este último no se «exige» una adecuación total, y permiten ordenar los resultados por relevancia, lo que al usuario le supone poder seleccionar. Dentro de la equiparación parcial podemos establecer dos grandes grupos: los que comparan con documentos individuales, y que basan esa equiparación en una serie de parámetros, y los que establecen la comparación con los miles de documentos virtuales en la web.

En los modelos en red podemos diferenciar, aparte de las ya mencionadas, las técnicas basadas en *clustering*, *browsing* y las que utilizan herramientas de visualización (o activación) de esa agrupación mediante representaciones gráficas) añadiendo relaciones entre los documentos. Estas últimas serán tratadas en un trabajo próximo.

4. MEDIDAS DE RELEVANCIA Y SIMILITUD

El conjunto de documentos en el que se puede buscar debe tener una forma específica de organizarse. Sin ella, la recuperación de muchos documentos llegaría a ser demasiado costosa. El conjunto organizado de documentos se denomina espacio documental (*document space*). En algunos modelos la estructura de la pregunta o perfil dificulta su inclusión en el espacio documental. Es el caso de la recuperación booleana. La pregunta aparece como una expresión lógica donde los propios términos son extraídos directamente del texto. En esta situación, la recuperación puede verse como una representación desde el espacio documental dentro del espacio de la pregunta. Cada documento es transformado en una representación compatible con la de la búsqueda. El sistema, entonces determina en dónde el documento transformado satisface los requisitos de la pregunta.

Si la pregunta no está incluida en el espacio documental, hay otra forma de verlo: como una función de evaluación sobre el espacio documental. En el caso más simple, ésto es una función característica, o lo que es lo mismo, que tiene el valor 1 si pertenece a los documentos relevantes a la pregunta, y 0 si se incluye en los no relevantes. En modelos más complejos la función se representa dentro de un rango de valores entre 0 y 1, cuyo intervalo representa los grados de relevancia frente a una pregunta.

Cuando la representación de la pregunta y el documento son similares, la pregunta puede considerarse como un puntero en el espacio documental. En cada caso, la imagen mental es aquélla en la que los documentos relevantes están agrupados cerca del puntero generado por la pregunta («cerca» debe ser interpretado en términos de pertenencia cercana en la distancia o en términos de encontrarse en dirección similar a la del origen del puntero en el espacio documental). La función de evaluación define un contorno que separa los documentos relevantes de los irrelevantes, o un conjunto de contornos correspondientes a los documentos de creciente relevancia con respecto a la búsqueda.

Estos modelos mentales, aplicados a la tarea automática de recuperar documentos en respuesta a una pregunta, pueden reducir el proceso a una evaluación de cada

documento sobre la base de alguna función computable o **medida**. La determinación que hay que tomar es dónde cada documento es relevante a una pregunta dada, ésto es, dónde es apropiado recuperar un documento en respuesta a una necesidad de información expresada. Como la relevancia está, en último extremo, en la mente del usuario, es difícil medirla directamente. Los sistemas de recuperación de la información deben primero medir lo que le sea posible al sistema sobre la base de las representaciones de los documentos y las preguntas. La mayoría de los sistemas identifican e igualan relevancia, en forma de equiparación entre el tema de la pregunta y el de la respuesta (*topicality*), con similitud léxica. Ésto puede argumentarse incluso cuando el proceso de recuperación se realiza con sofisticadas técnicas semánticas y otros principios de inteligencia artificial, en los que la base sigue siendo la similitud léxica, o la coincidencia o equiparación de palabras.

Se han propuesto muchas medidas de similitud diferentes (tal y como veremos en los siguientes apartados). La mayoría calculan como valor, como intervalo de proximidad (entre 0 y 1), con la interpretación de que un elevado valor en el intervalo (1) representa gran similitud y un valor bajo (0) representa la menor similitud. A este respecto, Korfhage dice que no hay ninguna razón para elegir esa escala[5].

Por otra parte, Belkin y Croft[6] también mantienen que una posibilidad a tener en cuenta es la utilización de «modelos extendidos». Así, no sólo se puede trabajar añadiendo pesos al modelo booleano convirtiéndolo en el booleano extendido, sino que también se puede utilizar las relaciones establecidas en este modelo para generar el vectorial extendido y el probabilístico extendido. Salton propuso la extensión del booleano a través de los pesos[7], y Croft[8] la del probabilístico[9].

5. COINCIDENCIA VS. EQUIPARACIÓN EN EL MODELO BOOLEANO

Es el primero y el más ampliamente adoptado, y se usa preferentemente en los sistemas comerciales. Está basado en la Lógica de Boole y en la teoría de conjuntos. Salton lo llama «sistema de ficheros invertidos»[10]. Es un sistema relativamente fácil de utilizar para los usuarios. Parte de la base de que uno/s documento/s son buscados por los usuarios a través de una serie de preguntas que se conciben con un conjunto de términos. La recuperación se basa en obtener los documentos que contengan esos términos. Utiliza los conocidos símbolos del álgebra de Boole («Y»,

---

[5] KORFHAGE, R.R. *Information Storage and Retrieval*. John Wiley, 1997. p. 81.

[6] BELKIN, N. & CROFT, B. Retrieval Techniques. *Annual Review of Information Science and Technology*, 1987, 22, p. 109-145.

[7] SALTON, G.; FOX, E.A. & WU, H. Extended Boolean information retrieval. *Comunications of the ACM*, 1983, 26, p. 1022-1036.

[8] CROFT, B. Boolean Queries and term Dependencies in Probabilistic Models of Document Retrieval Without Relevance Information. *Journal of Documentation*, 1986, 35, 4, p. 285-295.

[9] CROFT, W.B. Boolean Queries and Term Dependencies in Probabilistic Retrieval Models. *Journal of American Sosciety for Information Science* (en adelante, *JASIS*), 1986, 37,2, p. 71-77.

[10] SALTON, G. & MCGILL, M. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983, p. 24.

«O», «NO»), permite situar paréntesis para ordenar por orden de importancia, y también incluye los operadores de adyacencia («ADJ»).

La recuperación booleana empezó con las tarjetas perforadas. En ella cada pregunta es una función lógica de unas palabras dadas, que no son un documento en el sentido convencional. Como no hay similitud estructural entre documento y pregunta, ésta se considera como una entidad aparte, y la recuperación, con respecto a una pregunta dada, se entiende como una función definida en el espacio documental.

Las dificultades y límites de este modelo han sido reconocidas al poco tiempo de su puesta en marcha (Verhoeff, Goffman y Belzer[11]). Lancaster dice que «el uso del álgebra de Boole para preguntas en sistema de recuperación de la información automatizadas ha sido un error…»[12].

Un sistema de recuperación de la información booleano puro no suministra las bases para el desarrollo de los juicios de similitud significativos. Por definición, un documento dado satisface o no una cuestión planteada. Entonces, la representación definida por una pregunta es una función característica que divide el espacio documental en dos subconjuntos: los que coinciden y los que no.

Varias modificaciones han permitido establecer grados en el conjunto de documentos recuperados. Considerando la pregunta formulada en los siguientes términos:

A O B O C

ésta puede satisfacerse por los documentos que contengan al menos uno de los tres términos. De ellos, algunos pueden contener sólo uno de los términos, mientras otros pueden contener dos o los tres. Así, el conjunto recuperado puede estar clasificado por cuantos documentos contienen cada uno de esos tres términos, e incluso por términos específicos, separando así los documentos que tienen los términos A y B pero no C, de aquéllos que tienen A y C pero no B, y estos dos separados de los que contienen los tres términos. Este sistema permite generar juicios de proximidad para establecer clasificaciones dentro del conjunto de los documentos recuperados. Desde que los sistemas booleanos operan sobre la base de la presencia o ausencia de términos, muchos de los sistemas no incluyen el dato de la frecuencia de los términos. Aquí, la organización del conjunto recuperado no se puede basar en la medida de similitud dependiendo de la frecuencia. De cualquier forma, algunos de los interfaces visuales de sistemas de recuperación de información, incluyendo *VIBE* e *InfoCrytsal*, suministran información separada de los documentos de acuerdo con los criterios de la lógica de Boole. Y algunos motores de búsqueda como, por ejemplo, *Lycos*, trabajaba con este tipo de escalas.

---

[11] VERHOEFF, J.; COFFMAN, W. & BELZER, J. Inefficiency of the use of Boolean functions for information retrieval systems. *Communications of the ACM*, 1961, 4, p. 557-558, 594.

[12] LANCASTER, F. W. *Evaluation of on-line searching in Medlars (AIMTWX) by biomedical practitioners*. Report No. 101. Graduate School of Library Science, University of Illinois, Urbana, Illinois, 1972. Citado por Korfhage, p. 81.

De este modelo inicial han surgido variaciones como el «boeleano extendido», que asigna pesos a los términos de la búsqueda, o el *fuzzy*, que ha incluido la lógica difusa en sus postulados.

6. LA EQUIPARACIÓN EN EL ESPACIO VECTORIAL: INDICADORES

Al ser tan claras las deficiencias del modelo booleano, empezaron a generarse modelos alternativos de RI. Uno de los primeros fue el sistema *SMART*, desarrollado inicialmente en la Universidad de Harvard basado en el modelo vectorial (Lesk y Salton[13]). El sistema continuó desarrollándose por Salton y sus discípulos en la Universidad de Cornell, y ha mantenido su fuerza vital con experimentos que se realizan hasta hoy día.

Cuando se usa el modelo de recuperación vectorial, las medidas de similitud se asocian con la idea de distancia, siguiendo la filosofía de que los documentos cercanos en el espacio vectorial son altamente similares, o con una medida angular, basada en la tesis de que los documentos «en la misma dirección» están estrechamente relacionados. El sistema *SMART* usó inicialmente una medida angular.

La distancia de un documento con respecto a él mismo es 0. A este respecto, las medidas espaciales no son directamente adecuadas para medir la similitud en los cuales, valores altos, representan estrecha similitud documental. De cualquier forma, se pueden aplicar medidas de transformación en las cuales altos valores representen alta similitud (nótese que esto no es necesario si aceptamos que 0 es la medida de similitud máxima).

Una de las formas más utilizadas es la medida del coseno (Rigsbergen[14] Wilkilson et al[15]), que no mide la distancia, sino que desarrolla el coseno del ángulo entre los vectores que representan al documento y a la pregunta (o dos documentos, para medidas de similitud documental).

Para medir la similitud existen diferentes fórmulas[16]:

a) Producto de Dot, que es la coincidencia simple o intersección. Si *Qi* y *Dk* tienen un conjunto de términos comunes, éstos se pueden hallar a través de la intersección de ambos conjuntos.
b) Fórmula del coseno, aplicada tanto a la similitud pregunta/documento, como entre documentos:

---

[13] SALTON, G. *A flexible automatic system for the organization, storage, and retrieval of language data (SMART)*. Report ISR-5, sec. 1. Harvard Computation Laboratory, Cambridge, Massachusetts, 1964; SALTON, G. ed. *The SMART retrieval system. Experiments in automatic document processing*. Englewood Cliffs, New Jersey: Prentice Hall 1971, y LESK, M.E. *The SMART automatic text processing and document retrieval system*. Report ISR-8, sec. II. Harvard Computation Laboratory, Cambridge, Massachusetts. 1964.

[14] RIJSBERGEN, C. J. van. *Information retrieval*. 2nd. ed. London: Butterworths. 1979.

[15] WILKINSON, R. & HINGSTON, P. Using the cosine measure in a neural network for document retrieval. En *Proceedings of the 14th Annual international ACM/SIGIR Conference on Research and Development in Information Retrieval*. Chicago, 1991, p. 202-210.

[16] DOMINICH, S. *Mathematical Foundations of Information Retrieval*. Dordrecht (Holanda): Kluwer, 2001, p. 103-104.

$$\cos(DOC_i, Q_j) = \frac{\sum_{k=1}^{t} (TERM_{lk} \cdot QTERM_{jk})}{\sqrt{\sum_{k=1}^{t} (TERM_{ik})^2 \cdot \sum_{k=1}^{t} (QTERM_{jk})^2}}$$

c) Coeficiente de Dice:

$$s_{ik} = d_{ik} = 2 \frac{(D_i \quad Q_k)}{|D_i| + |Q_j|}$$

d) Coeficiene de Jaccard:

$$J_{ik} = \frac{(D_i \quad Q_k)}{|D_i| - |Q_k|}$$

e) Coeficiente de solapación:

$$O_{ik} = \frac{|D_i \quad Q_k|}{min(|D_i|, |Q_k|)}$$

Las medidas de distancia y angulares representan dos tipos diferentes de aproximación a los juicios de similitud. Las medidas de distancia son intrínsecas, basadas solamente en el grupo de documentos bajo consideración, mientras que las del ángulo del coseno son extrínsecas, y representan la visión del espacio documental desde un punto fijo, el origen. Si este punto se cambia, el ángulo entre documentos y preguntas puede variar también. Lo que significa que una medida angular no considera la distancia de cada documento respecto al origen, sino solamente la dirección. Aquí, los documentos que tienen el mismo vector de origen se consideran idénticos, a pesar del hecho de que puedan estar lejos en el espacio documental.

La medida del coseno (u otras angulares similares), en efecto, proyectan el espacio documental íntegro en una esfera *n*-dimensional de radio fijo alrededor del origen, donde *n* es la «dimensionalidad» del espacio documental, normalmente relacionado con el volumen del vocabulario. Así, elimina la distinción entre D_1, que está algo cercano al conjunto de términos, y D_2, que está mucho más cercano al conjunto de términos. En la práctica, la distancia y la medida angular tienen más o menos la misma calidad de resultados.

Por acuerdo se asigna 0 a un término que no está en una pregunta. Pero puede ser que no aparezca porque el indizador no pensara que fuera significativo más que porque no debiera aparecer. Exactamente igual puede pasar con el usuario, que no lo especifique porque piense que no es significativo para representar su pregunta. Así que, el valor 0 designa a los términos que no contienen el tema del documento y a los que, aunque sí deberían estar, no aparecen por alguna razón. El valor de cualquier término no específicamente incluido o excluido de la pregunta debería dejar

de ser indefinido. De cualquier forma, si ésto es un hecho, una pregunta (o un documento) es un subconjunto de dimensiones muy altas en el espacio.

Otro problema del modelo de espacio vectorial se establece con la relación entre los términos de un documento o pregunta. Por una parte, una de las mayores desventajas de este modelo es la habilidad para poner en práctica resultados conocidos del álgebra vectorial y cálculos del vector. Muchos de esos resultados, de cualquier forma, están basados en tomar el espacio vectorial definido como un conjunto de independencia lineal de vectores básicos. En el contexto documental, ésto puede significar definir un documento por significados de términos que no tienen relación entre ellos. Por otro lado, los términos que se usan para definir el modelo vectorial de RI no son claramente independientes: un ordenador digital es mucho más común que un «perro» digital (*tamagochi*). Lo que determina que hay una fuerte y estrecha relación entre los términos digital y ordenador. Algunas investigaciones han definido un conjunto de vectores básicos linealmente independientes para este modelo vectorial, pero los resultados no han sido ampliamente aceptados [Wong & Ziarko[17], Raghavan & Wong[18], Wong et al[19]].

Una de las desventajas del modelo booleno es que no incorpora los pesos, mientras que la del vectorial es que no es capaz de utilizar los conectores para generar expresiones lógicas. Así, nació el modelo booleano extendido. A las preguntas en este modelo extendido se le asignan pesos entre 0.0 y 1.0 a los términos. Considerando la pregunta expresada así:

$$A_{w1} * B_{w2}$$

en la que A y B son dos términos de la pregunta, w_1 *y* w_2 son los pesos y el «*» representa un conector u operador booleano.

7. EQUIPARACIÓN PROBABILÍSTICA

Los modelos anteriores están basados en la equiparación en la forma más «dura». En el booleano es o no coincidente, y en el vectorial el umbral de similitud es un conjunto, y si un documento no está no es similar y, por lo tanto, no recuperable.

La equiparación probabilística se basa en que, dados un documento y una pregunta, es posible calcular la probabilidad de que ese documento sea relevante para esa pregunta. Fue introducido por Maron y Kuhns[20], y más recientemente por Rijs-

---

[17] WONG, S.K.M. & ZIARKO, W. On generalized vector space model in information retrieval. *Annals of the Society of Mathematics of Poland, Series IV: Fundamentals of Information*, 1985, 8, 2, p. 253-267.

[18] RAGHAVAN, V. V.& WONG, S.K.M. A critical analysis of vector space model for information retrieval. *JASIS,* 1986, 37, 5, p. 279-287.

[19] WONG, S.K.M., ZIARKO, W. et al. On modelling of information retrieval concepts in vector spaces. *ACM Transactions on Database Systems*, 1987, 12, 2, p. 299-321.

[20] MARON, M.E.& KUHNS, J.L. On relevance, probabilistic indexing and information retrieval. *Journal of ACM*, 1960, 7, 3, p. 216-244.

bergen[21], Robertson[22], Bookstein[23], Kwok[24], Losee[25], Fuhr[26], Lee & Kantor[27], Wong y Yao[28], Cooper[29], Cooper[30], Gey[31]; Robertson & Waiker[32], Sebastiani[33], y otros.

Discutir la equiparación probabilística requiere algunas explicaciones sobre teoría de probabilidad. Asumir que en un momento dado podemos utilizar cualquier respuesta a una pregunta. Así, todas las probabilidades discutidas se toman en el contexto de esa pregunta. Si asumimos ésto, para el propósito de la discusión, el número de documentos de la base de datos que son relevantes a la pregunta son

---

[21] RIJSBERGEN, C.J. van. *Information Retrieval.* 2nd. ed. London: Betherworth, 1979.

[22] ROBERTSON, S. E., MARON, M.E. & COOPER, W.C. Probability of relevance: A unification of two competing models for document retrieval. *Information Technology: Research and Development,* 1982, 1, 1, p. 121.

[23] BOOKSTEIN, A. Outline of a general probabilistic retrieval model. Journal of Documentation. 1983, 39, p. 63-72; BOOKSTEIN, A. Probability and fuzzy set applications to information retrieval. *Annual Review of Information Science and Technology.* 1985, 20, p.117-152.

[24] KWOK, K.L. A document-document similarity measure based on cited titles and probability theory, and its application to relevance feedback retrieval. En *Research and Development in Information Retrieval: Proceedings of the Third Joint British Computer Society/ACM Symposium*, Cambridge, England, 1984, p. 221-231; KWOK, K.L. Experiments with cited titles for automatic document indexing and similarity measure in a probabilistic context. En *Proceedings of the Eighth Annual international ACM/SIGIR Conference on Research and Development in Information Retrieval*, Montreal, 1985, p.165-178; KWOK, K.L. A neural network for probabilistic information retrieval. En *Proceedings of the 12th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Cambridge, Massachusetts, 1989, p. 21-30.; KWOK, K.L. Experiments with a component theory of probabilistic information retrieval based on single terms as document components. *ACM Transactions on Information Systems*, 1990, 8, p. 363-386.

[25] LOSEE, R. M., Jr. Probabilistic retrieval and coordination level matching. *JASIS* 1987, 38, 4, p. 239-244.

[26] FUHR, N. Models for retrieval with probabilistic indexing. *Information Processing and Management*, 1989, 25, p. 55-72; FUHR, N. Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems*, 1989, 7, 3, p. 183-204.

[27] LEE, J. J. & KANTOR, P. A study of probabilistic information retrieval systems in the case of inconsistent expert judgments. *JASIS*, 1991, 42, 3, p. 166-172.

[28] WONG, S.K.M. & YAO, Y.Y. A probabilistic inference model for information retrieval. *Information Systems*, 1991, 16, 3, p. 301-321.

[29] COOPER, W. S., GEY, F.C. & DABNEY, D.P. Probabilistic retrieval based on staged logistic regression. En *Proceedings of the 15th Animal International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, 1992, p. 198-210; COOPER, W. S., GEY, F.C. & CHEN, A. Probabilistic retrieval in the TIPSTER collections: An application of staged Logistic regression. En *Overview of the First Text Retrieval Conference (TREC-1)*, ed. Donna K. Harman. Washington, D.C.: NIST Special Publication 500-207, 1993, p. 73-88.

[30] COOPER, W.S. The formalism of probability theory in IR: A foundation for an encumbrance? En *Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Dublin, 1994, p. 242-247.

[31] GEY, F. C. Inferring probability of relevance using the method of logistic regression. En *Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Dublin, 1994, p. 222-231; GEY, F. C. Evaluation of probabilistic retrieval methods. *Poster abstract in Proceedings of the 18th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, 1995, p. 370.

[32] ROBERTSON, S. E. & WAIKER, S. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. En *Proceedings of the 17th Annual international ACM/SIGIR Conference on Research and Development in information Retrieval*, Dublin, 1994, p. 232-241.

[33] SEBASTIANI, F. A probabilistic terminological logic for modelling information retrieval. En *Proceedings of the 17th Annual international ACM/SIGIR Conference on Research and Development in Information Retrieval*, Dublin, 1994. p. 122-130.

conocidos. Si un documento es seleccionado aleatoriamente de la base de datos hay cierta probabilidad de que sea relevante a la pregunta. Si una base de datos contiene *N* documentos, *n* de ellos son relevantes, entonces la probabilidad se estima en:

$$P(rel) = \frac{n}{N}$$

En concordancia con la teoría de la probabilidad, la de que un documento no sea relevante a una pregunta dada viene expresada por la siguiente formula:

$$P(\downarrow rel) = 1 - P(rel) = \frac{N - n}{N}$$

Obviamente, los documentos no son elegidos aleatoriamente, sino que se eligen sobre la base de la equiparación con la pregunta —basado en el análisis de los términos contenidos en ambos—. Así, la idea de relevancia está relacionada con los términos de la pregunta que aparecen en el documento.

Una pregunta dada divide la colección de documentos en dos conjuntos: los que responden a la pregunta y los que no. Sin embargo, todos los documentos seleccionados no son realmente relevantes. Entonces, debemos considerar la posibilidad de que un documento sea relevante o no, dado que haya sido ya seleccionado. Supongamos que un conjunto de documentos *S* de la base de datos ha sido seleccionado en respuesta a una pregunta. La cuestión es hasta qué punto éste es el conjunto que debería haber sido seleccionado en respuesta a la pregunta. Un criterio debe ser seleccionar el conjunto si es más probable que un documento del conjunto sea más relevante que otro que no lo es.

Evidentemente, la recuperación probabilística envuelve muchos cálculos y premisas. Numerosos experimentos demuestran que los procedimientos de recuperación probabilística obtienen buenos resultados. De cualquier forma, los resultados no son mucho mejores que los obtenidos en el modelo booleano y en el vectorial. Posiblemente en el nuevo contexto de la recuperación a texto completo de bases de datos heterogéneas en Internet, compliquen lo suficiente la recuperación como para que las técnicas de recuperación probabilística se utilicen más. Un ejemplo de ello es el trabajo de Gövert, Lalmas y Fuhr[34] que utilizan el enfoque probabilístico para facilitar la catogorización de documentos en la web.

También el trabajo de Picard y Savoy[35], que como se puede observar en la tabla siguiente otorga un porcentaje mayor de efectividad en RI al modelo probabilístico frente al tradicional de TREC y al más novedoso de similitud por enlaces.

---

[34] GÖVERT, N.; LALMAS, M. & FUHR, N. A probabilistic description-oriented approach for categorising Web Documents. *Proceedings of the 9th International Conference on Information and Knowledge Management*, 1999, p. 445-482.

[35] PICARD, J. & SAVOY, J. *Searching and Classyfing the web using hyperlinks: A logical approach* [Web]. Neuchatel: III, Universitè, 2001. [14-04-2004]. Disponible en <http://www.unine.ch/info/Gi/Papers/ictai.pdf>.

Tabla 1.
Comparación de resultados obtenidos de efectividad entre los experimentos TREC y el modelo probabilístico en la web

TRECeval	Probabilístico
0,253	0,267

Fuente: PICARD, J & SAVOY, J. *Searching and Classyfing the web using hyperlinks: A logical approach* [Web]. Neuchatel: III, Universitè, 2001. [04-04-2004]. Disponible en <http://www.unine.ch/info/Gi/Papers/ictai.pdf>.

8. EQUIPARACIÓN *FUZZY*

El problema de la equiparación anterior es evidentemente la estimación de la probabilidad. La equiparación difusa es similar, pero cambia la necesidad de estimar la probabilidad por una necesidad de primar la creencia sobre la relevancia de un documento dado. Ésta es una regla estricta que gobierna el uso de las probabilidades no aplicadas. En sentido real, un documento *fuzzy* no existe. Es decir, los autores no asignan grados de pertenencia a los términos o a los conceptos en sus documentos.

De cualquier forma, puede hacerse un «juicio difuso» sobre cuándo un documento debería estar en el conjunto de coincidentes con la pregunta. Ésta es la base del conjunto de términos que describen un documento o los términos usados en él. En la equiparación probabilística, el cálculo último devuelto sobre la probabilidad de que los términos de los documentos sean potencialmente relevantes a una pregunta, está contenida en los documentos relevantes y en los no relevantes. En la equiparación difusa, el cálculo se define basándose en el grado de pertenencia de los términos. La cuestión llega a ser tal, que el grado de confianza de que un documento contenga un término dado es relevante. Si ésto se usa para definir el grado de pertenencia, entonces este grado con respecto al conjunto de documentos relevantes, puede ser computado para cualquiera de los documentos.

El concepto de lógica difusa ha tenido diferentes enfoques. Si consideramos términos relacionados semánticamente, entonces podemos considerar en qué grado un término relacionado es equiparable con un término dado. Por ejemplo, en la pregunta sobre el «cocker spaniel», un documento que contiene el término «springer spaniel» no será equiparable, pero las dos razas de perros están relacionadas estrechamente, y en grado suficiente para que el documento pueda contener información útil. Otro documento sobre canes en general también puede contener información útil, pero quizás menos que el anterior. Así que, todo depende de cómo de específica sea la pregunta sobre el «cocker spaniel», o el juicio *fuzzy*.

Otro enfoque se basa en la búsqueda de descriptores que ofrezcan alguna indicación del valor de la información en el documento. Y éstos pueden ser tanto cualitativos como cuantitativos (Kamel et al[36]). Los indicadores cualitativos pueden ser

[36] KAMEL, M.; HADFIELD, B. & ISMAIL, M. Fuzzy query processing using clustering techniques. *Information Processing and Management* 1990, 26, 2, p. 279-293.

adjetivos calificativos como «pequeño», «grande», etc., que puedan indicar una escala cercana a la numérica u otros que no como, por ejemplo, «bonito», «feo», «colorido», etc. Los descriptores que pueden considerarse «cuantitativos» incluyen palabras como «pocos», «la mayoría de», etc. El uso de un término en un documento también puede ser descrito de manera *fuzzy*: tales como «muy significativo», «muy importante», y los resultados de la recuperación descrita como altamente relevantes o parcialmente relevantes. El problema, entonces, es decidir cómo cada término se traduce o transforma en una función de pertenencia asociada con la recuperación difusa. Un proceso de equiparación difusa puede aparecer combinada con otros procesos como, por ejemplo, el enfoque booleano extendido.

9. EQUIPARACIÓN EN EL ENFOQUE COGNITIVO

El enfoque cognitivo empezó a finales de los años 70, viéndose influenciado por la ciencia cognitiva, la creación de la revista *Cognitive Journal*, la contribución epistemológica de De Mey sobre el punto de vista cognitivo general[37], y sus relaciones con la Biblioteconomía y Documentación[38]. Pronto pasó a utilizarse en el desarrollo de la investigación en RI centrada en el usuario. Este punto de vista llegó a convertirse en alternativa claramente a la tradicional investigación en RI. Revisiones y discusiones del enfoque cognitivo durante los años 80 los podemos encontrar en la visión de Belkin[39], los ensayos críticos de Ellis[40], y el libro de Ingwersen[41].

Durante los primeros, la RI cognitiva se basó en las construcciones de De Mey[42], centradas principalmente en la famosa frase «cualquier proceso de información, sea perceptivo o simbólico, se ve mediatizado por un sistema de categorías o conceptos que, para el mecanismo de procesamiento de la información es un modelo de su mundo»[43]. A ésto, Ingwersen añade que el modelo del mundo consiste en estructu-

---

[37] DE MEY, MARC. The Cognitive Viewpoint: Its Development and Its Scope. En: CC77. *International Workshop on the Cognitive Viewpoint*; Ghent, Bel1977 March 24-26; Ghent, Belgium, University of Ghent, 1977; DE MEY, M. *The cognitive paradigm: an integrated understanding of scientific development*. Dordreicht: Reidel, 1982.

[38] DE MEY, M. The relevance of the cognitive paradigm for information science. En HABRO, O y KAJBERG, L. (ed.). *Theory and Application of Information Research. Proceedings of the 2nd. Internation Research Forum in Information Science*. London: Mansell, 1980, p. 49-61.

[39] BELKIN, N.J. The Cognitive Viewpoint in Information Science. *Journal of Information Science*, 1990, 16(1), p. 11-15.

[40] ELLIS, DAVID. A Behavioural Approach to Information Retrieval System Design. *Journal of Documentation*, 1989 September, 45(3),p. 171-212; ELLIS, D. The physical and cognitive paradigm in information retrieval research. *Journal of Documentation*, 1992, 48, p. 45-64.

[41] INGWERSEN, P. *Information Retrieval Interaction*. London: Taylor Graham, 1992.

[42] DE MEY, M. The relevance of the cognitive paradigm for information science. En HABRO, O y KAJBERG, L. (ed.). *Theory and Application of Information Research. Proceedings of the 2nd. Internation Research Forum in Information Science*. London: Mansell, 1980, p. 49-61.

[43] DE MEY, MARC. The Cognitive Viewpoint: Its Development and Its Scope. En: CC77. *International Workshop on the Cognitive Viewpoint*; Ghent, Bel1977 March 24-26; Ghent, Belgium, University of Ghent, 1977. p. xvi-xvii.

ra cognitivas que están determinadas por las «experiencias individuales y sociales/colectivas, educación, formación, etc»[44].

El estado cognitivo actual del individuo se ve afectado por el pasado y el contexto social. Este enfoque se vio inspirado en la sicología cognitiva rusa y las teorías de Luria. En su trabajo empírico sobre la clasificación humana de los objetos, Luria[45] demostró cómo el bagaje educativo, tales como la rutina de trabajo y situaciones provocan la forma humana de clasificar objetos y separar sus relaciones en categorías situacionales y categóricas (genéricas y parte-todo). Naturalmente, cada comportamiento clasificatorio impacta la percepción por individuos (o usuarios), por ejemplo, de la organización del conocimiento en bibliotecas.

Este enfoque ha generado varios intentos de unificación en un sólo modelo, pero hasta el momento los resultados obtenidos han sido subenfoques muy específicos, tales como el modelo global de polirepresentación de Ingwersen, el episódico de Belkin, el estratificado de Saracevic, o el de retroalimentación interactiva de Spink. Todos, haciendo hincapié y centrándose en algunos de los procesos cognitivos que pueden afectar al usuario desde que surge el denominado «estado anómalo de conocimiento» (o ASK) hasta que recibe la información a su pregunta, proponen la creación de un interfaz o dispositivo más amplio que los existentes, que detecte la necesidad informativa del usuario y le ayude a conceptualizarla. Pues bien, aquí es esta interfaz la que se encarga de que la equiparación sea la mejor para que pregunta y respuesta no vayan paralelas, sino que consigan converger en la información que el usuario necesita. La intervención de un intermediario con más valor que las interfaces de los sistemas tradicionales, deja en manos de sus diseñadores el tipo de equiparación que se debe establecer, dejando paso a la que consideren mejor. Motivo por el cual desde el paradigma cognitivo no hay una apuesta por ningún tipo de equiparación concreta.

10. CONCLUSIÓN

Tras el estudio se ha demostrado que no es lo mismo la coincidencia (o equiparación) total del modelo booleano, que la equiparación parcial del modelo vectorial basado en la función de similitud, el probabilístico en la regla bayesiana y el difuso en la lógica *fuzzy*. Sin embargo, todas son aproximaciones de cómo intentar lograr que la recuperación sea más efectiva, sin que ninguna de momento haya aparecido como la mejor solución. Posiblemente sea porque no exista una solución global, porque dependiendo de los sistemas, las preguntas, los usuarios o la naturaleza de los documentos unos sean más válidos que otros.

Si bien la relevancia puede ir unida a conceptos conocidos como similitud semántica o probabilidad, tampoco debemos olvidar que actualmente están apareciendo nuevas propuestas, que aunque todavía no están integradas en modelos com-

---

[44] INGWERSEN, P. Search Porcedures in the Library. *Journal of Documentation,* 1982, 38, 3, p. 168.

[45] LURIA, A.R. *Cognitive development: its Cultural and Social Foundations*. Cambridge (Ma): Harvard University Press, 1976. 175 p.

pactos, son en sí soluciones posibles y probables, por no decir alternativas, a los tradicionales enfoques. Se trata de la recuperación de información a través del análisis de las citas (o de las «sitas» en Internet), de las co-palabras (o palabras que sin ser términos compuestos suelen ir unidas en los textos de los documentos), la adyacencia o vecindad de las mismas (como el ejemplo del «ordenador digital» expuesto en el modelo vectorial), entre otras.

BIBLIOGRAFÍA

BATES, M.J. Where should the person stop and information search start?. *Information Processing and Management*, 1990, núm. 26, p. 575-591.

BELKIN, N. J., KANTOR, P.; FOX, E. A. & SHAW, J. A. Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management*, 1995, núm. 31, p. 431-448.

BELKIN, N.J. ODDY, R.N. & BROOKS, H.M. ASK for Information Retrieval: part I. Background and theory. *Journal of Documentation*, 1982, 38, p. 61-71.

— ASK for Information Retrieval: part II. Result of a design study. *Journal of Documentation*, 1982, 38, p. 145-164.

BELKIN, N.J. The Cognitive Viewpoint in Information Science. *Journal of Information Science*, 1990, 16(1), p. 11-15.

BOLLMAN-SDORA, P. & RAGHAVAN, V.V. On the delusiveness of adopting a common space for modelling IR objects: Are queries documents?. *Journal of American Society for Information Science*, 1993, 44, 10, p. 579-587.

BORGMAN, C.L. All users of information retrieval systems are not created equal: an exploration into individual differences. *Information Processing and Management*, 1989, p. 237-251.

CAID, W.R.; DUMAIS, S.T. y GALLANT, S.L. Learned vector-space models for document retrieval. *Information Processing and Management*, 1995, p. 419-429.

CORDÓN GARCIA, O.; MOYA ANEGÓN, F. y ZARCO, C. A GA-P algorithm to automatically formulate extended boolean queries for a fuzzy information retrieval system by means of GA-P techniques. *Mathware & Soft Computing*, 2000, 7, p. 309-322.

DOMINICH, S. *Mathematical Foundations of Information Retrieval*. Dordrecht (Holanda): Kluwer, 2001.

EGGHE, L.& ROUSSEAU, R. Duality in Information Retrieval and the Hypergeometric Distribution. *Journal of Documentation*, 1997, 53, 5, p. 488-496.

FABA PEREZ, C.; GUERRERO BOTE, V. & MOYA ANEGÓN, F. Methods for analysing web citations: a study of web-coupling in a closed environment. *Libri*, 2004, (en prensa).

FIDEL, R. Online searching styles: a case-study-based model of searching behaviour. *Journal of American Society for Information Science*, 1984, 35, p. 211-221.

— Searchers' selection of search keys: I. The selection routine. II. Controlled vocabulary or free-text searching. III. Searching styles. *Journal of American Society for Information Science*, 1991, 42, 7, p. 490-527.

FUHR, N. Probabilistic Models in Information Retrieval. *The Computer Journal*, 1992, vol. 35, n.º 3, p. 243-255.

INGWERSEN, P. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation*, 1996, núm. 52, p. 3-50.

KORFHAGE, R. *Information Storage and Retrieval*. New York: John Wiley & Sons, 1997.

LOSEE, R.M. Term Dependence: A Basis for Luhn and Zipf Models. *Journal of American Society for Information Science*, 2001, 52, 12, p. 1019-1025.

MOYA ANEGÓN, F. Sistemas avanzados de recuperación de la información. En: LÓPEZ YEPES, J. (ed.). *Manual de Ciencias de la Documentación*. Madrid: Pirámide, 2002, p. 553-600.

RIJSBERGEN, C.J. van. A theoretical basis for the use of coocurrence data in Information Retrieval. *Journal of Documentation*, 1977, 13, p. 106-119.

— *Information Retrieval.* 2nd. ed. London: Betherworth, 1979.

SALTON, G. *Automatic Information Organisation and Retrieval.* New York: McGraw-Hill, 1966.

SALTON, G. & LESK, M.E. Computer Evaluation of indexing and text processing. *Journal of the ACM*, 1968, 15, 1, p. 8-36.

SALTON, G. (ed.). *The SMART Retrieval Systems-Experimental in Automatic Document Processing*. Englewood Cliffs (NJ): Prentice Hall, 1971, p. 324-336.

SALTON, G. & McGill, M. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.