Introducción a los modelos clásicos de Recuperación de Información

Fidel CACHEDA

Dpto. de Tecnologías de la Información y las Comunicaciones. Universidad A Coruña
fidel@udc.es

RESUMEN
En este artículo se sintetizan las principales características de los modelos clásicos de Recuperación de Información (RI). Por una parte, el modelo booleano constituye el más simple de estos modelos, lo que conlleva que la calidad de sus resultados puede ser mejorada sensiblemente. El modelo probabilístico establece un modelo teórico fundamental dentro del campo de la RI basado en la teoría de probabilidades, intentando interpretar toda la incertidumbre que rodea el proceso de RI. El modelo vectorial se basa en considerar a los documentos (y las consultas) como vectores de términos y calcular su similitud en un espacio de n dimensiones.

Palabras clave: Recuperación de información, modelos clásicos.

Introduction to the Classic Models of Information Retrieval

ABSTRACT
In this article main features of the classic models of information retrieval are smmarized. On the one hand, Boolean model is the simplest of these models, which implies that the quality of their results can be improved significantly. The probabilistc model provides a essential theoretical model in the filed of information retrieval, based on probability theory, trying to interpret all the uncertainty surrounding the process of IR. The vectorial model considers a document (an the query) as vectors of terms and calculate its similarity in an space of dimensions.

Keywords: information retrieval, classic models.

1. INTRODUCCIÓN

La Recuperación de Información (RI) es un ámbito que ha tomado una gran importancia en la última década. Esto está relacionado directamente con toda la infor-

mación disponible en la World Wide Web y la necesidad de herramientas que nos permitan gestionar, recuperar y filtrar esta información.

CONSULTA

En primer lugar, es interesante destacar la diferencia entre Recuperación de Datos (RD) y Recuperación de Información. En la RD la información disponible se encuentra archivada Documentos siguiendo una estructura clara (p.e. en forma de ordenados tabla), para posteriormente poder recuperar exactamente lo que se está buscando. La calidad de la RD se suele medir en función de la Figura 1. Un sistema de RI velocidad de búsqueda y del espacio de usuario y de cada situación. En este sentido, la almacenamiento requerido. Un ejemplo típico de calidad de la RI se mide en función de su RD sería el que nos ofrece cualquier base de velocidad y espacio de almacenamiento, pero en datos. mayor en medida en función de la calidad de la información recuperada para el usuario.

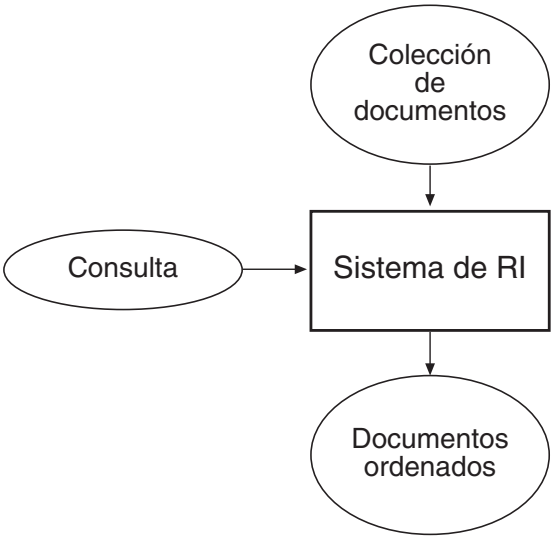

Figura 1

La RI se caracteriza por su indefinición. Los documentos están formados por un texto que no Según Ricardo Baeza-Yates, el problema al que tiene una estructura clara ni es sencillo crearla, y se enfrenta la RI se puede definir como: Dada lo mismo sucede con las necesidades de una necesidad de información (definida a partir información de los usuarios. de una consulta, el perfil del usuario, etc.) y una colección de documentos, se trata de presentar al …

En RI se dice que no existe la respuesta correcta usuario, de manera ordenada, un subconjunto de para una consulta, sino que habrá documentos los documentos más relevantes (Figura 1). más o menos relevantes en función de cada …

Esto se puede subdividir en dos subproblemas, estudiándose y evaluándose cada uno de ellos de manera independiente:

— Definir un modelo para calcular la relevancia de un documento frente a una necesidad de información (parte en la que se centra este artículo).

— Diseñar los algoritmos y las estructuras de datos para realizarlo de manera eficiente.

Dentro de un sistema de RI, el proceso de RI se divide en dos partes: el proceso de indexación y el proceso de búsqueda. Estos dos procesos están unidos a través del índice (Figura 2).

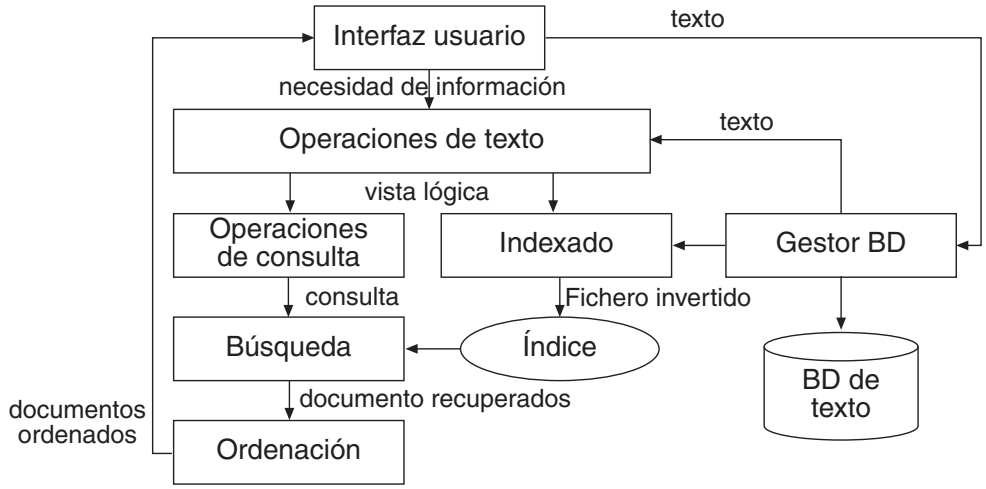

Figura 2. El proceso de RI.

Inicialmente, se realizan una serie de operaciones sobre el texto de los documentos que forman parte de la colección, con el objetivo de crear una vista homogénea de los mismos (p.e. conversiones a minúsculas, acentuación, etc.).

El proceso de indexado se encarga de construir el índice asociado al texto. El índice es una estructura de datos crítica ya que permitirá el acceso eficiente a grandes volúmenes de datos. Este proceso de indexación es muy costoso, desde el punto de vista del tiempo empleado y el espacio de almacenamiento requerido.

A partir de la indexación de los documentos, se inicia el proceso de recuperación o de búsqueda. Para ello, el usuario especificará una necesidad de información que será analizada y transformada siguiendo las mismas operaciones realizadas sobre el texto de los documentos.

A continuación, se pueden aplicar ciertas operaciones sobre la consulta (p.e. la expansión de la consulta) antes de procesar la consulta y extraer los documentos más relevantes de la colección. Este procesamiento debe ser realizado con unos tiempos de respuesta mínimos, al utilizar como base la estructura del índice construida previamente. Antes de ser enviados los documentos recuperados al usuario, son ordenados de acuerdo a un criterio de relevancia.

Los modelos de RI establecen el mecanismo empleado para realizar el procesamiento de las consultas de los usuarios. En lo que resta de este artículo se describen las principales características de los modelos tradicionales de RI.

2. MODELOS DE RI

Un modelo no es más que una representación abstracta de un proceso, probablemente del mundo real. En el día a día se utilizan habitualmente múltiples modelos para estudiar las propiedades de un proceso, intentar extraer conclusiones y, en la mayoría de los casos, hacer predicciones. Es de suponer que, cuánto más se ajuste el modelo a la realidad, mejores serán estas predicciones.

Los modelos de predicción meteorológica constituyen uno de los ejemplos más utilizados hoy en día, aunque existen otros muchos: modelos financieros, juegos de simulación, etc.

En el caso de la RI, un modelo de RI es una representación abstracta del proceso de RI. Es decir, a partir de una necesidad de información y una colección de documentos, el modelo intentará predecir si un documento puede ser considerado relevante (o no), y en que grado. El objetivo final de un modelo de RI es obtener una ordenación para los documentos relevantes para esa necesidad de información.

Básicamente, los modelos de RI deben encontrar una representación para:

— Los documentos
— Las consultas
— Las funciones de ordenación de los resultados.

Un documento se representará normalmente como un conjunto de términos o palabras clave. Un término deberá ser una palabra que permita identificar el contenido o la temática del documento.

Dado que no todas las palabras parecen tener un mismo peso a la hora de dar sentido a un documento, para representar un documento se pueden aplicar ciertas técnicas como:

— La eliminación de las palabras comunes (p.e. artículos, preposiciones, etc.) que, en la mayoría de los casos, no aportan nada significativo a un documento.
— Extraer la estructura de un documento y dar un mayor peso a aquellos términos considerados relevantes. Esta técnica es utilizada por los buscadores en Internet, y consiste en darle un peso mayor a los términos que forman el título de la página o aparecen resaltados de alguna manera (p.e. en negrita).

De la misma manera se puede representar una consulta. Si consideramos una consulta como un documento de pequeño tamaño, las mismas consideraciones realizadas sobre la representación de un documento pueden ser aplicadas a la representación de una consulta.

De manera más formal, se puede definir un modelo de RI como una cuádrupla [*D,Q,F,R(qi,dj)*], donde:

— *D*: es el conjunto de representaciones de los documentos
— *Q*: es el conjunto de representaciones de necesidades de información de los usuarios
— *F*: es el marco de modelado de documentos, consultas y sus relaciones
— *R(qi,dj)*: es la función de ranking que define el orden de los documentos, respecto a la consulta

En RI existen tres modelos clásicos de RI, proporcionando cada uno de ellos una visión diferente del proceso de RI:

— Modelo booleano: basado en la teoría de conjuntos.
— Modelo probabilístico: basado en la teoría de probabilidades.
— Modelo vectorial: basado en el álgebra.

2.1. MODELO BOOLEANO

El modelo booleano se basa en la teoría de conjuntos, teniendo en cuenta que la relevancia de un documento es binaria: un documento será relevante para una consulta o totalmente irrelevante.

Un documento se representa como un conjunto de términos, de tal forma que un término estará presente o ausente de un determinado documento, sin contemplar la posibilidad de establecer diferentes grados de pertenencia.

Las consultas se expresan mediante expresiones booleanas que se corresponden con operaciones sobre conjuntos (Figura 3):

— AND: intersección de conjuntos.
— OR: unión de conjuntos.
— NOT: complementario de un conjunto.

El resultado obtenido será un conjunto de documentos (por lo tanto, sin ordenar) con aquellos documentos que satisfagan la expresión booleana de la consulta.

Las principales ventajas del modelo booleano se centran en su sencillez. Esto hace que este modelo sea muy intuitivo, especialmente para aquellos usuarios más expertos, y fácil de implementar y formalizar. Esto motivó que fuese el modelo elegido en los primeros sistemas de RI.

Las principales desventajas de este modelo se centran en su excesiva rigidez. No es posible ordenar los resultados obtenidos y tampoco se tienen en cuenta el número de cláusulas verificadas en una consulta de tipo OR. No se tiene en cuenta el número de veces que aparece una palabra en un documento y las consultas booleanas pueden resultar confusas para aquellos usuarios menos expertos.

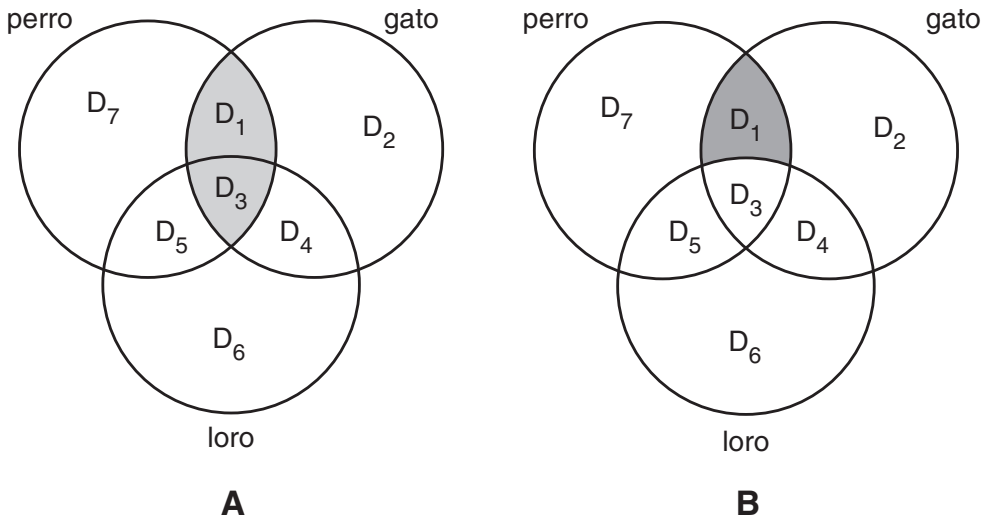

Figura 3. *Consultas del modelo booleano: perro AND gato (a), perro AND gato AND NOT loro (b) Se consideran los documentos D1-D7 formados por los correspondientes términos: D1 = (perro AND gato), D2 = (gato), D3 = (perro AND gato AND loro), D4 = (loro AND gato), D5 = (loro AND perro), D6 = (loro), D7 = (perro).*

2.2. MODELO PROBABILÍSTICO

El modelo probabilístico fue formulado por Stephen Robertson y Sparck Jones en 1977. Este modelo se basa en que en el proceso de RI es intrínsicamente impreciso.

Dentro del propio proceso, hay determinados aspectos que son no deterministas, por ejemplo:

— La representación que hace una consulta de la necesidad de información del usuario.

— La representación de los documentos en el sistema.

Teniendo esto en cuenta, el modelo probabilístico postula que la mejora manera de poder representar esto es mediante la teoría de probabilidades.

Este modelo intenta estimar la probabilidad de que, dada una consulta *q*, un documento *d* sea relevante para esa consulta. Esto se denota como: *P(Rel | d)*. En el modelo se intenta obtener un conjunto de documentos relevantes (denominado *R*), que deberá maximizar la probabilidad de relevancia.

Un documento se considera relevante si su probabilidad de ser relevante, *P(Rel | d)*, es mayor que la probabilidad de no ser relevante, *P(noRel | d)*. Dicho de otra manera, para calcular la similitud de un documento con una consulta, *sim(q, d)*, se calcula la división entre ambas probabilidades:

$$_{(,)} = P(\text{Re } l \mid d)$$
$$simd\ q$$
$$(\text{Re } l \mid d)$$
$$Pno$$

Aplicando el Teorema de Bayes y tras una serie de simplificaciones, se aproxima el valor de la similitud mediante la siguiente expresión:

$$_{(,)} = (\mid Re\ )\ \boldsymbol{l}$$
$$\boldsymbol{Pd} simd\ q$$
$$(\mid no\ \text{Re})$$
$$Pd\ l$$

Donde *P(d | Rel)* representa la probabilidad de que, sabiendo que hemos seleccionado un documento relevante, ese documento sea *d*. Respectivamente, *P(d | noRel)* representa la probabilidad análoga de que *d* no sea relevante.

Teniendo en cuenta que la probabilidad de relevancia de un documento es estimada a partir de las probabilidades de los términos que lo componen, podemos calcular el grado de similitud de un documento con una consulta.

El modelo probabilístico se basa en un proceso iterativo. Este proceso se inicia con un primer conjunto de documentos relevantes, que es paulatinamente recalculado en función de la información que proporciona el usuario de aquellos documentos que considera relevantes y no relevantes.

La principal ventaja de este modelo consiste en que constituye un modelo teórico importante que permite representar el proceso de RI. Además, el conjunto resultante proporciona una ordenación de los documentos en base a su probabilidad de relevancia.

Dentro de sus desventajas, cabe destacar la necesidad de iniciar el modelo a partir de una primera estimación del conjunto de documentos relevantes, y el hecho de que no se tiene en cuenta el número de veces que cada término aparece en un documento a la hora de estimar su probabilidad de relevancia.

2. MODELO VECTORIAL

En el modelo vectorial los documentos se representan como un vector de términos, y viceversa. Las consultas se modelan como un vector de términos y el modelo recupera los documentos relevantes en función de la similitud de los vectores de los documentos con el vector de la consulta, en un espacio n-dimensional.

En una primera aproximación, la similitud entre un documento *d* y una consulta *q* se puede medir como el producto interno de *q* y *d*:

$$simd\ q\ (,) = \sum_{i=1}^{n} q_i \times d_i$$

Donde *qi* y *di* son los valores de las posiciones iésimas de los vectores *q* y *d*, respectivamente. En otras palabras, consiste en contar el número de términos comunes entre el documento y la consulta.

El producto interno presenta la limitación de que no considera la longitud de los documentos. Esto hace que aquellos documentos más largos, y que probablemente contengan un mayor número de términos de la consulta, tenga una mayor probabilidad de ser seleccionados como relevantes. Para solucionar esto se suele normalizar el producto interno dividiendo entre la longitud del documento (definido como el número de términos).

De manera más formal, la similitud en el modelo vectorial se corresponde con el ángulo entre el vector del documento y el vector de la consulta. Si el ángulo entre ambos vectores es 0º son idénticos, mientras que si el ángulo es de 90º no tienen absolutamente nada en común.

Por lo tanto, la medida base que se utiliza para medir la similitud es la denominada distancia coseno:

$$(qd)$$
$$\cdot$$
$$\sum ii$$
$$-1\ -1$$

$$(,) = \sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2} \quad i=1 (simd\ q = q \cdot q)\ (\bullet d \cdot d)$$

$$n\ \Sigma q_i^2 \cdot \Sigma d_i^2 \quad i=1\ i=1$$

Donde el numerador no es más que el producto interno del vector de documentos y el vector de consulta, y los términos del denominador simplemente son los factores de normalización de la longitud de la consulta y del documento.

En esta primera aproximación, el peso de los términos en los vectores de documentos y consultas es binario (presencia o ausencia). Esto plantea dos inconvenientes:

— No se tiene en cuenta la frecuencia de un término en un documento.

— Se considera que todos los términos son igual de importantes, cuando no es así. Por ejemplo, en la consulta "el perro", el término "perro" es mucho más significativo que el término "el".

Para solucionar esto se incorpora el *modelo tf-idf* a la hora de asignar un peso a los términos:

— En el vector del documento se almacena la frecuencia del término en el documento (componente *tf*: *term frequency*).
— Para valorar aquellos términos más significativos se les da más peso a los términos que ocurren en un menor número de documentos (componente *idf*: *inverse document frequency*).

La componente *tf* se calcula directamente como la frecuencia de un término en un documento.

La componente *idf* se calcula como *idfi = log(N/ni)*. Donde *N* representa el total de documentos en la colección y *ni* representa el número de documentos en donde aparece el término i-ésimo, todo ello suavizado mediante la función logarítmica.

Finalmente, el peso de un término se calcula como el producto de la componente *tf* por la componente *idf*.

Las principales ventajas del modelo vectorial son las siguientes:

— Permite aciertos parciales, ya que un documento puede ser considerado relevante aunque no incluya todos los términos de la consulta.
— La ordenación de los resultados se realiza en base a varios factores: frecuencia de los términos, importancia de los términos y sin primar a los documentos más largos.
— Además, permite una implementación eficiente para grandes colecciones de documentos.

Por otra parte, sus principales desventajas se centran en que se pierde parte de la información sintáctica y semántica del documento y que se basa en la independencia de los términos dentro de un documento, aseveración que no siempre se mantiene.

3. CONCLUSIONES

En este trabajo se han expuesto de manera resumida los modelos clásicos de RI. Desde un punto de vista más pragmático, resulta interesante plantearse cuál de estos modelos es el más eficaz considerando la calidad de los resultados que proporciona. Como era de prever, el modelo booleano ofrece los resultados más modestos en este sentido. Por otra parte, el modelo probabilístico y el modelo vectorial proporcionan valores equiparables en cuanto a la calidad de sus resultados.

Estos modelos constituyen la base teórica sobre la que se han desarrollado extensiones y ampliaciones que intentan presentar nuevas aportaciones al mundo de la RI. Dentro de estas extensiones simplemente mencionar algunas de ellas:

— Extensiones al modelo booleano: modelo booleano extendido y modelo de conjuntos difusos.
— Extensiones al modelo probabilístico: Okapi BM25 y Divergence from Randomness.
— Extensiones al modelo vectorial: modelo vectorial generalizado, Latent Semantic Indexing (LSI) y redes neuronales.

BIBLIOGRAFÍA

[1] Modern Information Retrieval, R. Baeza-Yates y B. Ribeiro-Neto. Addison-Wesley, 1999.
[2] Information Retrieval, K. van Rijsbergen. http://www.dcs.gla.ac.uk/Keith/Preface.htm.