Motores de Búsqueda Web Tarea Tema 3: Limitaciones de la recuperación de información tradicional en la Web

José Alberto Benítez Andrades

71454586A

Motores de Búsqueda Web
Máster en Lenguajes y Sistemas Informáticos - Tecnologías del Lenguaje en la Web
UNED
10/02/2011

10 de febrero de 2011

Tarea Tema 3: Limitaciones de la recuperación de información tradicional en la Web

Enunciado del ejercicio

Se trata de hacer un pequeño informe (1-3 páginas) señalando cuáles son las limitaciones de los modelos y técnicas desarrollados en Recuperación de Información tradicional a la hora de buscar en la Web.

Resolución

1. Introducción

En este trabajo explicaré 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. Y finalmente 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.

2. Modelo Booleano

En qué consiste

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:

— 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.

Desventajas o limitaciones

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. Otra desventaja es que el modelo no diferencia entre los operadores AND y OR y las palabras del lenguaje natural 'and' y 'or'.

3. Modelo Probabilístico.

En qué consiste

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 intrínsecamente 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: $P(Rel \mid d)$. 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, $P(Rel \mid d)$, es mayor que la probabilidad de no ser relevante, $P(noRel \mid d)$. Dicho de otra manera, para calcular la similitud de un documento con una consulta, $sim(q, d)$, se calcula la división entre ambas probabilidades:

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.

Desventajas o limitaciones

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.

4. Modelo Vectorial

En qué consiste

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*:

$$simd\ q\ (,) = \Sigma\ qi \times di$$
$$i=1$$

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:

$$(qd)$$
$$\Sigma\ ii$$
$$\vdash 1\ -1$$

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

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.

Desventajas o limitaciones

El principal inconveniente del modelo de espacio vectorial es que de ninguna manera lo suscriben los valores de las componentes del vector que deben ser. Los primeros experimentos de Salton (1971) ya sugerían que el plazo de ponderación no es un problema trivial.

Un segundo inconveniente del modelo de espacio vectorial es que no es posible incluir las dependencias de plazo en el modelo, por ejemplo para el modelado de las frases o términos adyacentes. Sin embargo, es posible dar una interpretación geométrica de las consultas booleanas estructurado.

Un tercer problema con el modelo de espacio vectorial es su aplicación. El cálculo de la medida del coseno requiere las necesidades de los valores de todos los componentes del vector, pero estos no están disponibles en una arquitectura de archivo invertido. En la práctica, se deben utilizar los valores normalizados y el algoritmo vectorial del producto. Cualquiera de los pesos normalizados tiene que ser almacenado en el archivo invertido, o la normalización de los valores que tienen que ser almacenados por separado.

Ambos tienen mucho más espacio de almacenamiento que serían necesarios para el modelo booleano.

Referencias

1. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.39.531&rep=rep1&type=pdf

2. http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/maldonado_n_mf/capitulo2.pdf

3. http://comminfo.rutgers.edu/~aspoerri/InfoCrystal/Ch_2.html

4. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.2128&rep=rep1&type=pdf