RECUPERACIÓN Y ORGANIZACIÓN DE LA INFORMACIÓN

INGENIERÍA INFORMÁTICA

RECUPERACIÓN Y ACCESO A LA INFORMACIÓN

MODELOS DE RECUPERACION

AUTOR:

Rubén García Broncano NIA 100065530 grupo 81

INDICE

1- INTRODUCCIÓN A LOS MODELOS DE RECUPERACIÓN.

Según la mayoría de estudios que se han estado realizando en los últimos años la **recuperación y organización de la información** es uno de los aspectos que han cobrado un mayor relevancia. En la actualidad estos estudios resaltan la vital importancia que ha cobrado ese campo. Esto se debe en gran medida a que los buscadores de internet están situados como el primer método utilizado para obtener cualquier tipo de información sea para el uso que sea (académico, lúdico, empresarial).

Debido a esto es de vital importancia conocer cuales son los métodos o **modelos de recuperación** utilizados por los grandes buscadores (**booleano, probabilístico, vectorial**). En los últimos años y debido a los intereses económicos derivados de buenos posicionamientos en los distintos buscadores se está produciendo un boom en todos los campos relacionados con este tema, por tanto es necesario conocer como se estructuran los modelos de recuperación con anterioridad.

La principal *clasificación para los modelos de recuperación de información* es la siguiente:

- **Modelos clásicos:** Entre los que se encuentran los modelos probabilístico, booleano y vectorial, los cuales describimos con todo detalle en este documento.
- **Modelos estructurales:** Entre los que destacan listas no sobrepuestas y el método de los nodos proximales.

Esta será la clasificación que utilizaremos en nuestro ya que es la mas extendida y a nuestro parecer la de mayor utilidad.

2.- CLASIFICACIÓN LOS MODELOS DE RECUPERACION.

Para facilitar el seguimiento de los contenidos, esta sección esta estructurada en varias subsecciones, en este punto realizamos una breve descripción de las principales características de los distintos modelos, para después en la pertinente subsección realizar un análisis completo.

- **Modelo Booleano:** Es un modelo de recuperación simple, basado en la teoría de conjuntos y el álgebra booleana. Dada su inherente simplicidad y su pulcro formalismo ha recibido gran atención y sido adoptado por muchos de los primeros sistemas bibliográficos comerciales. Su estrategia de recuperación está basada en un criterio de decisión binario (pertinente o no pertinente) sin ninguna noción de escala de medida, sin noción de un emparejamiento parcial en las condiciones de la pregunta.

- **Modelo Vectorial:** El modelo de recuperación vectorial o de espacio vectorial propone un marco en el que es posible el emparejamiento parcial, asignando pesos no binarios a los términos índice de las preguntas y de los documentos. Estos pesos de los términos se usan para computar el grado de similitud entre cada documento guardado en el sistema y la pregunta del usuario.

- **Modelo Probabilistico:** El modelo de recuperación probabilístico se basa en la equiparación probabilística, dados un documento y una pregunta, es posible calcular la probabilidad de que ese documento sea relevante para esa pregunta.

3.- MODELO DE RECUPERACIÓN BOOLEANO

El **modelo de resuperación booleano** es uno de los métodos más utilizados para la recuperación de información. Este modelo se basa en la agrupación de documentos, los cuales están compuestos por conjuntos de términos y en la concepción de las preguntas como expresiones booleanas, de ahí deriva el nombre de ***modelo de recuperación booleano.*** La principal característica es la consideración de la relevancia como un carácter puramente binario. Dentro del modelo, se presenta el lenguaje de consulta, y el mecanismo de indización utilizando los denominados índices inversos o archivos fantasma.

3.1.-Características Principales

Es un modelo de recuperación simple, basado en la teoría de conjuntos y el álgebra booleana. Dada su inherente simplicidad y su pulcro formalismo ha recibido gran atención y sido adoptado por muchos de los primeros sistemas bibliográficos comerciales. Su estrategia de recuperación está basada en un criterio de decisión binario (pertinente o no pertinente) sin ninguna noción de escala de medida, sin noción de un emparejamiento parcial en las condiciones de la pregunta.

Para el **modelo de recuperación booleano**, las variables de peso de los términos índice son todas binarias. A pesar de estos inconvenientes, el modelo booleano es todavía el modelo dominante en los sistemas comerciales de bases de datos de documentos y proporciona un buen punto de partida.

En este modelo el método de representación como ya hemos mencionado es definir a los documentos como un conjunto de términos de indexación o palabras claves.

- **Diccionario:** Conjunto de todos los términos T = {t1, t2, t3, ...}.
- **Documento:** Conjunto de términos del diccionario donde tiene valor Di = {t1, t2, t3, ...} donde cada uno de los ti = Verdad si es una palabra clave del documento.

Las preguntas son expresiones booleanas cuyos componentes son términos de nuestro diccionario:

- **Operadores** : O (U), Y (∩), No (-)

El algoritmo utilizado en el **método booleano** nos permite calcular el valor de la función de semejanza. Como entrada tenemos dos listas ordenadas ascendentemente y como salida una lista ordenada con la mezcla de las dos listas de entrada.

El método de ordenación puede ser el número de identificación de los documentos que agrupan los términos a recuperar. Para todo esto necesitaremos, una función que nos devuelva los identificadores de los documentos que contienen el término de la búsqueda, lo cual es sencillo si miramos el archivo invertido y luego se mezclan las listas.

Los beneficios de utilizar este método es que es un modelo de recuperación sencillo. Mientras que la problemática es que básicamente tenemos que considerar la relevancia como un aspecto puramente binario, y las extensiones que se pueden especificar para el manejo de pesos en el **modelo booleano**.

4.- MODELO DE RECUPERACION VECTORIAL.

El **modelo de recuperación vectorial o de espacio vectorial** propone un marco en el que es posible el emparejamiento parcial a diferencia del modelo de recuperación booleano**,** asignando pesos no binarios a los términos índice de las preguntas y de los documentos. Estos pesos de los términos se usan para computar el grado de similitud entre cada documento guardado en el sistema y la pregunta del usuario.

4.1.- Características Generales.

Ordenando los documentos recuperados en orden decreciente a este grado de similitud, el **modelo de recuperación vectorial** toma en consideración documentos que sólo se emparejan parcialmente con la pregunta, así el conjunto de la respuesta con los documentos alineados es mucho más preciso (en el sentido que empareja mejor la necesidad de información del usuario) que el conjunto recuperado por el modelo booleano. Los rendimientos de alineación del conjunto de la respuesta son difíciles de mejorar.

La mayoría de los motores de búsqueda lo implementan como estructura de datos y que el alineamiento suele realizarse en función del parecido (o similitud) de la pregunta con los documentos almacenados.

4.2.- Funcionamiento.

La idea básica de este modelo de recuperación vectorial reside en la construcción de una matriz (*podría llamarse tabla*) de términos y documentos, donde las filas fueran estos últimos y las columnas correspondieran a los términos incluidos en ellos. Así, las filas de esta matriz (que en términos algebraicos se denominan ***vectores***) serían equivalentes a los documentos que se expresarían en función de las apariciones (***frecuencia***) de cada término. De esta manera, un documento podría expresarse de la manera:

- **d1=(1, 2, 0, 0, 0, ..., 1, 3) :** Siendo cada uno de estos valores el número de veces que aparece cada término en el documento.

La longitud del vector de documentos sería igual al total de términos de la matriz (el número de columnas).

De esta manera, un conjunto de m documentos se almacenaría en una matriz de m filas por n columnas, siendo n el total de términos almacenamos en ese conjunto de documentos. La segunda idea asociada a este modelo es calcular la similitud entre la pregunta (que se convertiría en el vector pregunta, expresado en función de la aparición de los n términos en la expresión de búsqueda) y los m vectores de documentos almacenados. Los más similares serían aquellos que deberían colocarse en los primeros lugares de la respuesta.

4.3.- Cálculo de la similitud.

Se dispone de varias fórmulas que nos permiten realizar este cálculo, la más conocida es la ***Función del Coseno***, que equivale a calcular el producto escalar de dos vectores de documentos ***(A y B)*** y dividirlo por la raíz cuadrada del sumatorio de los componentes del ***vector A*** multiplicada por la raíz cuadrada del sumatorio de los componentes del ***vector B***.

De esta manera se calcula este valor de similitud. Como es obvio, si no hay coincidencia alguna entre los componentes, la similitud de los vectores será cero ya que el producto escalar será cero (circunstancia muy frecuente en la realidad ya que los vectores llegan a tener miles de componentes y se da el caso de la no coincidencia con mayor frecuencia de lo que cabría pensar).

También es lógico imaginar que la **similitud máxima** sólo se da ***cuando todos los componentes de los vectores son iguales***, en este caso la función del coseno obtiene su máximo valor, la unidad. Lo normal es que los términos de las columnas de la matriz hayan sido filtrados (supresión de palabras vacías) y que en lugar de corresponder a palabras, equivalgan a su raíz *'stemmed'* (agrupamiento de términos en función de su base léxica común, por ejemplo: economista, económico, economía, económicamente, etc.). Generalmente las tildes y las mayúsculas/minúsculas son ignorados. Esto se hace para que las dimensiones de la matriz, de por sí considerablemente grandes no alcancen valores imposibles de gestionar.

No obstante podemos encontrar excepciones a la regla general, tal como parece ser el caso de **Yahoo!**, que no ignora las palabras vacías.

Para finalizar, la del coseno no es la única función de similitud. Existen otras, las cuales no son difíciles de calcular sino más bien de interpretar y que por tanto son menos aplicadas en Recuperación de Información.

4.4.- Modelo de Recuperación Vectorial Generalizado .

La idea del ***modelo generalizado*** es tomar el **grupo de vectores** *mi* que son **ortogonales** y adoptarlo como el conjunto de vectores bases para los subespacios de interés. Ortogonalidad no significa que las palabras índices son independientes. Por el contrario, las palabras índices son ahora correlacionadas por los vectores *mi* .

4.4.1- Funcionamiento.

La independencia de las palabras clave en un **modelo vectorial** implica que el conjunto de vectores es linealmente independiente. Frecuentemente esta linealidad es interpretada como que los vectores son ortogonales.

En el ***modelo vector generalizado***, los pesos (weights) son considerados independientes pero no ortogonales. Sea el conjunto de ***palabras índices*** { *k1, k2, ... kt* } y ***los pesos*** *wi,j* asociados a las ***palabras índices y documentos*** *[ki, d*j]. Si los pesos son binarios, toda posible concurrencia de palabras índices pueden ser representada por el conjunto de 2t "minterms" dados por *m1* = (0,0,...0), *m2* = (1,0,...0) y *mt* = (1,1,...1).

Considere la función gi*(m*j) que retorna el peso {0,1} de la palabra índice *ki* en el minterm *m*j. El minterm *m1* que contiene sólo 0 significa que el documento no tiene ninguna de las palabras índices y el minterm *mt* significa que el documento contiene todas las palabras índices.

5.- MODELO DE RECUPERACIÓN PROBABILISTICO .

Este tema presenta un modelo de recuperación clásico como es el **modelo de recuperación Probabilístico**, donde la base principal de su funcionamiento es el cálculo de la probabilidad de un documento de ser relevante a una pregunta dada. 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.

5.1.- Características Principales.

Dentro de la recuperación probabilística, utilizaremos el **modelo de recuperación probabilístico de independencia de términos binarios** donde: "*La probabilidad de los términos es independiente (un término es independiente de los otros)" y "los pesos asignados a los términos son binarios".*

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.

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) = 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(↓rel) = 1 - P (rel) = 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.

5.2.-Ventajas y desventajas de los modelos probabilísticos.

Numerosos experimentos demuestran que los procedimientos del modelo de recuperación probabilístico 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.

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, los **modelos de recuperación probabilísticos** envuelven muchos cálculos y premisas, que los expresados en este documento, pero como un acercamiento a este tema, en nuestra opinión es más que suficiente.