

UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA



Estrategias de planificación para motores de búsqueda verticales

Danilo Fernando Bustos Pérez

Profesor Guía: Dra. Carolina Bonacic Castro
Profesor Co-guía: Dr. Mauricio Marín Caihuán

Trabajo de Titulación presentado en conformidad
a los requisitos para obtener el Título de
Ingeniero Civil Informático

SANTIAGO DE CHILE
2013

© Danilo Fernando Bustos Pérez

Se autoriza la reproducción parcial o total de esta obra, con fines académicos, por cualquier forma, medio o procedimiento, siempre y cuando se incluya la cita bibliográfica del documento.

AGRADECIMIENTOS

Dedicado a

RESUMEN

Resumen en Castellano

Palabras Claves: keyword1, keyword2 .

ABSTRACT

Resumen en Inglés

Keywords: keyword1, keyword2 .

ÍNDICE DE CONTENIDOS

Índice de Figuras	iii
-------------------	-----

Índice de Tablas	iv
------------------	----

1. Introducción	1
------------------------	----------

1.1. Antecedentes y motivación	2
1.2. Descripción del problema	2
1.3. Objetivos y solución propuesta	2
1.3.1. Objetivo General	2
1.3.2. Objetivos Específicos	2
1.3.3. Alcances	2
1.3.4. Solución propuesta	2
1.3.5. Características de la solución	2
1.3.6. Propósito de la solución	2
1.4. Metodología y herramientas de desarrollo	2
1.4.1. Metodología	2
1.4.2. Herramientas de desarrollo	2
1.5. Resultados obtenidos	2
1.6. Organización del documento	2

2. Marco teórico	3
-------------------------	----------

2.1. Motores de búsqueda verticales	3
2.2. Índice invertido	5
2.3. Estrategias de evaluación de <i>queries</i>	6

<i>ÍNDICE DE CONTENIDOS</i>	ii
2.3.1. TAAT	6
2.3.2. DAAT	7
2.3.3. Consideraciones	7
2.4. Operaciones sobre listas invertidas	8
2.4.1. OR	8
2.4.2. AND	9
2.4.3. WAND	10
2.5. <i>Ranking</i>	11
2.5.1. TF-IDF	12
2.5.2. BM25	13
3. Conclusiones	14
Referencias	15

ÍNDICE DE FIGURAS

2.1. Arquitectura típica de un motor de búsqueda	4
2.2. Índice invertido	6
2.3. Operación OR	9
2.4. Operación AND	9
2.5. Operación AND	11

ÍNDICE DE TABLAS

CAPÍTULO 1. INTRODUCCIÓN

1.1 ANTECEDENTES Y MOTIVACIÓN

1.2 DESCRIPCIÓN DEL PROBLEMA

1.3 OBJETIVOS Y SOLUCIÓN PROPUESTA

1.3.1 Objetivo General

1.3.2 Objetivos Específicos

1.3.3 Alcances

CAPÍTULO 2. MARCO TEÓRICO

En este capítulo se exponen los conceptos teóricos del presente trabajo de tesis. Primero se explica qué es un motor de búsqueda vertical. Luego se definen las estrategias de evaluación de *queries*. Posteriormente se describen las diferentes operaciones sobre listas invertidas. Finalmente, se explica el concepto de *ranking*.

2.1 MOTORES DE BÚSQUEDA VERTICALES

A medida que pasa el tiempo y la Web sigue creciendo, los motores de búsqueda se convierten en una herramienta cada vez más importante para los usuarios. Estas máquinas ayudan a los usuarios a buscar contenido dentro de la Web, puesto que conocen en cuales documentos de la Web aparecen qué palabras. Si estas máquinas no existieran, los usuarios estarían obligados a conocer los localizadores de recursos uniformes (URL) de cada uno de los sitios a visitar. Además, los motores de búsquedas en cierto modo conectan la Web, ya que existe un gran número de páginas Web que no tienen referencia desde otras páginas, siendo el único modo de acceder a ellas a través de un motor de búsqueda.

Un motor de búsqueda está construido por diversos componentes. Su arquitectura típica se puede ver en la Figura 2.1. Existe un proceso denominado *crawling*, éste posee una tabla con los documentos Web iniciales en los que se extrae el contenido de cada uno de ellos. A medida que el *crawler* comienza a encontrar enlaces a otros documentos Web, la tabla de documentos a visitar crece. El contenido que se extrae en el procedimiento de *crawling* es enviado al proceso de

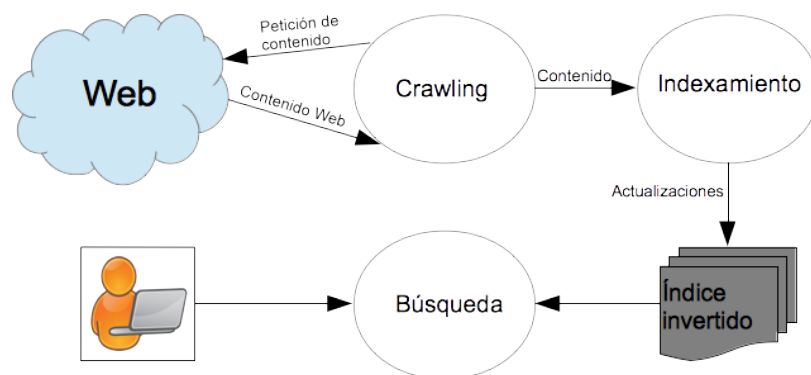


FIGURA 2.1: Arquitectura típica de un motor de búsqueda

indexamiento, este se encarga de crear un índice de los documentos ya visitados por el *crawler*.

Dado el volumen de datos involucrado en el procesamiento, se debe tener una estructura de datos que permita encontrar cuáles documentos contienen las palabras presentes en la búsqueda que llega al sistema. Todo esto dentro de un período de tiempo aceptable. El índice invertido (Zobel & Moffat, 2006) es una estructura de datos que contiene una lista con todas las palabras que el proceso de *crawling* ha visto. Asociado a cada palabra se tiene una lista de todos los documentos Web donde ésta palabra aparece mencionada. El motor de búsqueda construye esta estructura con el objetivo de acelerar el proceso de las búsquedas que llegan al sistema. El proceso de búsqueda es el encargado de recibir la consulta (*query*), generar un *ranking* de los documentos Web que contienen las palabras de la *query* y finalmente generar una respuesta. Las diversas formas de calcular la relevancia de un documento será explicado en secciones posteriores.

En un motor de búsqueda se pueden encontrar diversos servicios tales como (a) cálculo de los mejores documentos Web para una cierta *query*; (b) construcción de la página Web en la que se mostrará al usuario los resultados de la *query*; (c) publicidad relacionada con las *queries*; (e) sugerencia de *queries*; entre muchos otros servicios.

En los sistemas de recuperación de la información modernos como los motores de búsqueda, lo que se hace hoy en día es agrupar computadores para procesar una *query* y

obtener la respuesta para ésta. Este conjunto de computadores recibe el nombre de *cluster*.

La diferencia entre un motor de búsqueda vertical y uno general, es que el primero se centra solo en un contenido específico de la Web. El *crawler* debe extraer contenido solo de aquellas páginas Web que están dentro del dominio permitido. Al ser un dominio acotado, los documentos Web a procesar serán menos y por lo tanto, la lista de los términos del índice invertido serán eventualmente de menor tamaño. Sin embargo, en un motor de búsqueda vertical las actualizaciones al índice invertido ocurren con mayor frecuencia.

2.2 ÍNDICE INVERTIDO

Es una estructura de datos que contiene todos los términos (palabras) encontrados por el *crawler*. A cada uno de los términos, está asociado una lista de documentos (páginas Web) que contienen dicho término. Adicionalmente, se almacena información que permita realizar el *ranking* de documentos para generar la respuesta a las *queries* que llegan al sistema, por ejemplo, el número de veces que aparece el término en el documento. Esta lista recibe el nombre de lista invertida.

Para construir un índice invertido se debe procesar cada palabra que existe en un documento Web, registrando su posición y la cantidad de veces que éste se repite. Cuando se procesa el término con la información asociada correspondiente, se almacena en el índice invertido (ver Figura 2.2).

El tamaño del índice invertido crece rápido y eventualmente la memoria RAM se agotará antes de procesar toda la colección de documentos. Cuando la memoria RAM se agota, se almacena en disco el índice parcial hasta aquel momento, se libera la memoria y se continúa

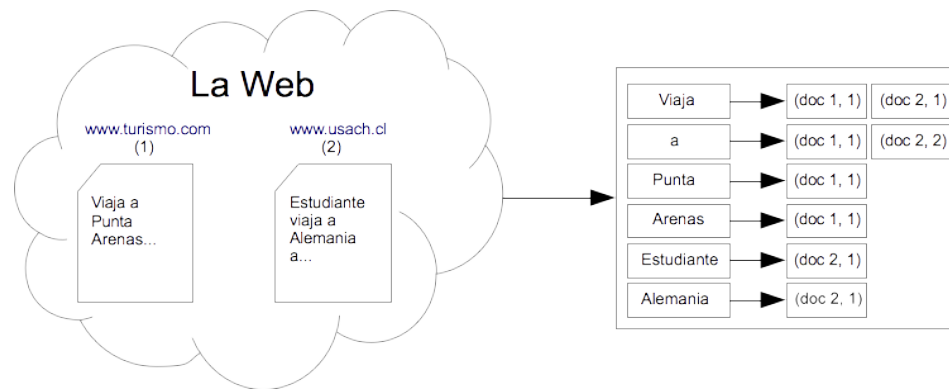


FIGURA 2.2: Índice invertido

con el proceso. Además, se debe hacer un *merge* de los índices parciales uniéndolos las listas invertidas de cada uno de los términos involucrados.

2.3 ESTRATEGIAS DE EVALUACIÓN DE *QUERIES*

Existen dos principales estrategias para encontrar los documentos y calcular sus respectivos puntajes de una determinada *query*. Estas son (a) *term-at-a-time* (TAAT) y (b) *document-at-a-time* (DAAT).

2.3.1 TAAT

Este tipo de estrategia procesa los términos de las *queries* uno a uno y acumula el puntaje parcial de los documentos. Las listas invertidas asociadas a un término son procesadas

secuencialmente, esto significa que los documentos presente en la lista invertida del término t_i , obtienen un puntaje parcial antes de comenzar el procesamiento del término t_{i+1} . La secuencialidad en este caso es con respecto a los términos contenidos en la *query*.

2.3.2 DAAT

En este tipo de estrategias se evalúa la contribución de todos los términos de la *query* con respecto a un documento antes de evaluar el siguiente documento. Las listas invertidas de cada término de la *query* son procesadas en paralelo, de modo que el puntaje del documento d_j se calcula considerando todos los términos de la *query* al mismo tiempo. Una vez que se obtiene el puntaje del documento d_j para la *query* completa, se procede al procesamiento del documento d_{j+1} .

2.3.3 Consideraciones

Cuando se tiene un índice invertido pequeño, las estrategias TAAT rinden adecuadamente, sin embargo cuando los índices invertidos son de gran tamaño las estrategias DAAT poseen dos grandes ventajas: (a) Requieren menor cantidad de memoria para su ejecución, ya que el puntaje parcial por documento no necesita ser guardado y (b) Explotan el paralismo de entrada y salida (I/O) más eficientemente procesando las listas invertidas en diferentes discos simultáneamente. Además existen técnicas para optimizar el proceso de las estrategias recientemente descritas

(Turtle & Flood, 1995).

2.4 OPERACIONES SOBRE LISTAS INVERTIDAS

Cuando una *query* llega al motor de búsqueda, cada término tiene asociado una lista con todos los documentos en los cuales aparece. El sistema debe decidir qué documentos se analizarán para obtener la respuesta con el conjunto de los K mejores. A continuación se presentan las diferentes formas de operar las listas invertidas de una *query*.

2.4.1 OR

Este operador toma las listas invertidas de cada uno de los términos de la *query* y ejecuta la disyunción entre ellas. El resultado de este operador es una lista invertida con todos los documentos que contengan al menos un término de la *query*. Finalmente, esta lista invertida se ocupará para obtener los mejores K documentos. Un simple ejemplo se muestra en la FIGURA 2.3.

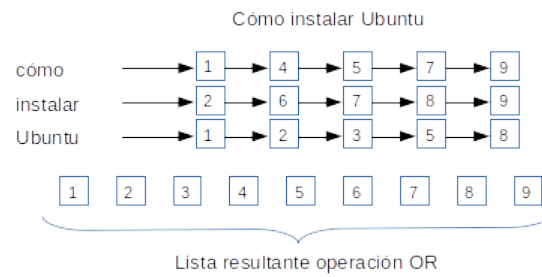


FIGURA 2.3: Operación OR

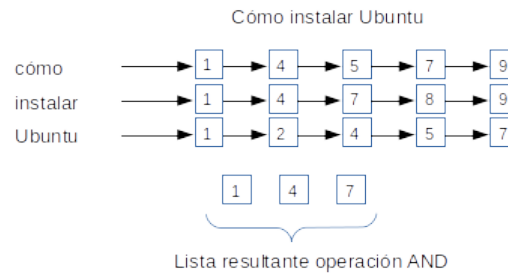


FIGURA 2.4: Operación AND

2.4.2 AND

Este operador ejecuta la conjunción entre las listas invertidas de los términos de una *query*. Se obtiene una lista invertida con los documentos que contengan todos los términos de la *query*. Se debe notar que aquí se obtiene una lista resultante de menor tamaño que la obtenida en el operador OR (Ver FIGURA 2.4).

2.4.3 WAND

Método de evaluación de *queries* para obtener eficientemente el conjunto de K documentos que mejor satisfacen una *query* dada. *WAND* (Broder et al., 2003) (*Weak AND*) es un proceso menos estricto que el método *AND* y es basado en dos niveles. Dentro del proceso de evaluación de una *query*, uno de los procesos más costoso en términos de tiempo es el proceso de *scoring*. Este proceso corresponde a entregarle a cada uno de los documentos analizados, un puntaje que representa la relevancia del documento para una *query* dada, esto se denomina evaluación completa o cálculo del puntaje exacto del documento.

El objetivo de *WAND* es minimizar la cantidad de evaluaciones completas de los documentos ejecutando un proceso de dos niveles. En el primer nivel se intenta omitir rápidamente grandes porciones de las listas invertida, lo que se traduce en ignorar el cálculo del puntaje exacto de grandes cantidades de documentos. El intento de omitir el cálculo completo de un documento es debido a que en motores de búsqueda a gran escala, este es un proceso que requiere de mucho tiempo para llevarse a cabo y depende de factores como la cantidad de ocurrencia del término dentro del documento, el tamaño del documento, entre otros.

En el primer nivel se itera sobre los documentos del índice invertido de cada término y se identifica candidatos usando una evaluación aproximada. En el segundo nivel, aquellos documentos candidatos son completamente evaluados y su puntaje exacto es calculado. De esta forma se obtiene el conjunto final de documentos.

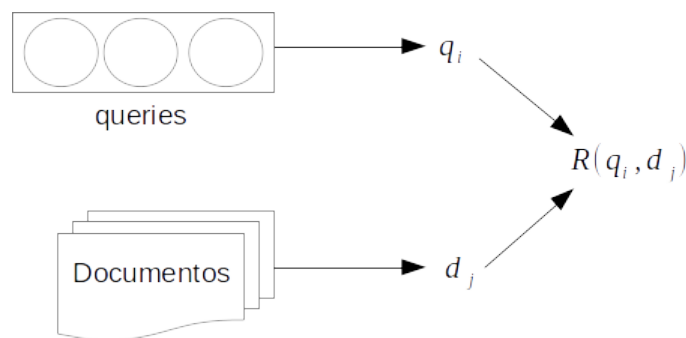


FIGURA 2.5: Operación AND

2.5 RANKING

Los sistemas de recuperación de información como los motores de búsqueda deben ejecutar un proceso el cual asigna un puntaje a documentos con respecto a una determinada *query*, este proceso se denomina *ranking* (Baeza-Yates & Ribeiro-Neto, 2011). Como se puede ver en la Figura 2.5, este proceso toma como entrada la representación de las *queries* y documentos, y asigna un puntaje (*score*) a un documento d_j dada una *query* q_i .

Un motor de búsqueda guarda billones de documentos que están formados por términos o palabras, estos términos no todos poseen la misma utilidad para describir el contenido del documento. Determinar la importancia de una palabra en un documento no es tarea sencilla, para ello se asocia un peso positivo $w_{i,j}$ a cada término t_i del documento d_j . De esta forma, para un término t_i que no aparezca en el documento d_j se tendrá $w_{i,j} = 0$. La asignación de pesos a los términos permite generar un *ranking* numérico para cada documento en la colección.

2.5.1 TF-IDF

El tf-idf (*term frequency - inverse document frequency*) es un estadístico que tiene por objetivo reflejar cuán importante es una palabra para un documento en una colección o corpus. Este estadístico se divide en dos partes, el primero corresponde a la frecuencia de la palabra o término en un documento (tf) y que en su versión más sencilla se utiliza la frecuencia bruta del término t en el documento d ($f(t,d)$). El segundo término corresponde a la frecuencia inversa de documento (idf) y se utiliza para observar si es que el término es común en el corpus. El idf obtiene calculando el logaritmo de la división entre el número total de documentos del corpus y el número de documentos que contienen el término. De esta forma se tiene:

$$tf(t, d) = \frac{f(t, d)}{\max f(w, d) : w \in d}$$

$$idf(t, D) = \log \frac{|D|}{1 + |d \in D : t \in d|}$$

$$tfidf(t, d, D) = tf(t, d) * idf(t, D)$$

Por lo que el estadístico *TF-IDF* incrementa proporcionalmente al número de veces que la palabra aparece en el documento, sin embargo es compensado por la frecuencia de la palabra en la colección completa de documentos o corpus. Notar que la compensación ayuda a controlar el hecho de que algunas palabras son generalmente más comunes que otras.

2.5.2 BM25

Es una función de *ranking* de documentos basada en los términos que aparecen en la *query* que llega al motor de búsqueda. *BM25* pertenece a una amplia gama de funciones de puntuación y está basada en los modelos probabilísticos de recuperación de la información (Baeza-Yates & Ribeiro-Neto, 2011).

Dada una *query* Q que contiene los términos q_1, \dots, q_n , el *ranking BM25* del documento D se calcula como:

$$score(D, Q) = \sum_{i=1}^n IDF(q_i) * \frac{f(q_i, D) * (k + 1)}{f(q_i, D) + k * (1 - b + b * \frac{|D|}{prom(docs)})}$$

donde $f(q_i, D)$ es la frecuencia en que aparece el término q_i en el documento D ; $|D|$ es el número de palabras o términos en el documento D ; $prom(docs)$ es la media de número de palabras de los documentos en el corpus; k y b son constantes que depende de las características del corpus en el que se está haciendo la búsqueda, por lo general se asignan los valores de $k = 2$ o $k = 1.2$ y $b = 0.75$; finalmente, $IDF(q_i)$ es la frecuencia inversa de documento para el término q_i .

CAPÍTULO 3. CONCLUSIONES

REFERENCIAS

Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (2011). Modern information retrieval - the concepts and technology behind search, second edition.

Broder, A. Z., Carmel, D., Herscovici, M., Soffer, A., & Zien, J. (2003). Efficient query evaluation using a two-level retrieval process. En *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, (pág. 426–434). New York, NY, USA: ACM.

URL <http://doi.acm.org/10.1145/956863.956944>

Turtle, H., & Flood, J. (1995). Query evaluation: Strategies and optimizations. *Inf. Process. Manage.*, 31(6), 831–850.

URL [http://dx.doi.org/10.1016/0306-4573\(95\)00020-H](http://dx.doi.org/10.1016/0306-4573(95)00020-H)

Zobel, J., & Moffat, A. (2006). Inverted files for text search engines. *ACM Comput. Surv.*, 38(2).

URL <http://doi.acm.org/10.1145/1132956.1132959>