

# Búsquedas en Internet

## Ingeniería de sistemas de información

DANIEL LÓPEZ GARCÍA  
LOTHAR SOTO PALMA  
*Universidad de Granada*  
2 de mayo de 2017

### Índice

<b>1. Introducción</b>	<b>2</b>
<b>2. La WWW</b>	<b>2</b>
<b>3. Modelo de arquitectura referente de motores de búsqueda</b>	<b>2</b>
3.1. Elementos del modelo . . . . .	2
3.2. Medida de relevancia . . . . .	3
<b>4. Modelo de arquitectura multi-capa para motores de búsqueda</b>	<b>4</b>
<b>5. Modelo de arquitectura del motor de búsqueda de Google</b>	<b>4</b>
5.1. Características . . . . .	4
5.2. Metodología . . . . .	4
5.3. PageRank . . . . .	4

### Índice de figuras

1. Modelo de referencia . . . . .	2
-----------------------------------	---

# 1. Introducción

Actualmente los motores de búsqueda son necesarios en nuestras vidas, cuando un usuario tiene una necesidad de información se dirige a una fuente de conocimiento para intentar suplirla, debido al reciente crecimiento del número de webs, Internet se ha convertido en una fuente de conocimiento masiva, como consecuencia los motores de búsqueda web se han vuelto muy importantes y la necesidad de obtener una arquitectura para estos sistemas que optimice la realización de búsquedas es ahora una necesidad. En este documento vamos a analizar cada uno de los elementos de la arquitectura de un motor de búsqueda.

Pero antes de eso es necesario conocer el entorno en el que trabajan dichos motores de búsqueda:

## 2. La WWW

La WWW o World Wide Web es un sistema de distribución de documentos de hipertexto que están interconectados y son accesibles vía internet. El usuario a través de un navegador web visualiza el contenido proporcionado por la web compuesto por elementos multimedia y texto.

Pero la web es un entorno con unas características un poco adversas a la hora de poder hacer búsquedas sobre ellas esto es debido a su escala, volatilidad y variabilidad de la calidad de las mismas.

En la web no hay un diseño general, aparte de incluir información verídica puede incluir información obsoleta y contradicciones, el contenido es generado dinámicamente en el tiempo por lo que es necesario volver a analizar cada web para incluirlas en los resultados de las búsquedas de forma rutinaria y por último tiene un gran crecimiento llegando a duplicarse cada pocos meses.

## 3. Modelo de arquitectura referente de motores de búsqueda

### 3.1. Elementos del modelo

Los motores de búsqueda que operan en internet tienen comunmente una arquitectura centralizada, los elementos que componen esta arquitectura son:

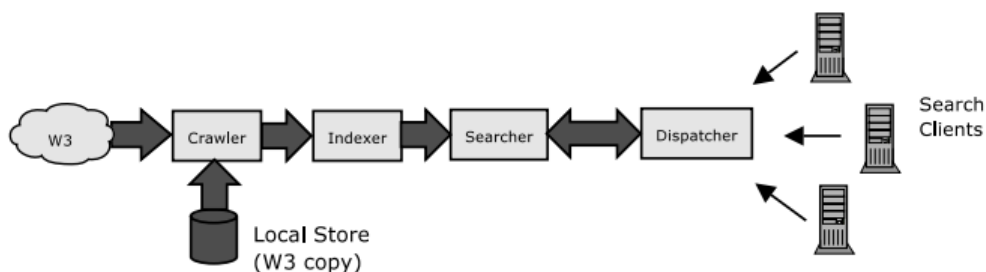


Figura 1: Modelo de referencia

- **Crawler:** Un crawler o una araña es normalmente un módulo que se encarga de la agregación al sistema de los documentos que posteriormente serán indexados para su búsqueda. De forma obvia la obtención de documentos de Internet no es una cuestión sencilla de resolver debido a la naturaleza descentralizada de la web. El funcionamiento básico de un crawler o araña es:

- Partimos de un conjunto semilla de URLs iniciales
- Establecemos una cola con prioridad en la que se irán incluir las URLs anteriores
- Para cada URL de la cola:
  - Descargamos el contenido de la web
  - Extraemos los términos de indexación
  - Incluimos los links de la web en la cola con prioridad
- Se vuelve a añadir a la cola la URL analizada

Pero este proceso básico no es lo único a tener en cuenta de hecho las arañas suelen necesitar la ayuda de los servidores webs que analizan, con la aportación de datos que permiten entre otras cosas excluir webs del servidor de la araña o moderar el número de peticiones por minuto.

Existen muchas heurísticas y algoritmos para el crawling y la mayoría de ellos se basan en el análisis de hipertexto en los documentos.

- **Indexer:** El indexador es el módulo que se encarga de a partir de los datos obtenidos con el proceso de crawling construir una estructura de acceso rápido (normalmente una tabla hash) llamada índice. Existen diversos modelos de indexación algunos de ellos son:
  - **Modelo binario:** La interpretación principal de este modelo consiste en la elaboración de un índice centrándose en la aparición o no de los términos en los documentos.
  - **Modelo Vectorial:** Este modelo a diferencia del anterior tiene en cuenta la frecuencia de aparición de un término en los documentos (tf), sin embargo a lo largo del tiempo han surgido modelos que tienen en cuenta la frecuencia inversa de aparición de un término en un documento (idf), y otro modelo que combina las dos ideas anteriores llamado modelo tf-idf.
  - **Modelo Probabilístico:** El modelo probabilístico se centra en la probabilidad de aparición de un término en un documento.
- **Searcher:** El buscador es aquel módulo que normalmente se encarga de el procesamiento de las consultas que se realizan normalmente a través de un entorno web con un navegador, es el encargado de la realización de consultas sobre el índice creado por el indexador anterior.
- **Dispatcher:** Por último el dispatcher es el encargado de la recepción de la consulta realizada desde el buscador y enviarla al buscador que realizará el cotejamiento con el índice para posteriormente obtener una lista ordenada por una puntuación denominada relevancia de URLs que será mostrada a los usuarios a través del cliente.

### 3.2. Medida de relevancia

El grado de relevancia de un documento web es una medida que resulta necesaria definir para que el resto de procedimientos puedan trabajar correctamente, sin embargo la definición de esta medida no es simple puesto que depende normalmente del modelo de indexación que se lleve a cabo. La relevancia de un documento normalmente depende directamente de la consulta que se ha realizado, pero hay algunos casos en los que se supone que la relevancia puede verse como la suma de dos componentes una que depende de la consulta llamada componente dinámica y otra que no depende llamada componente estática, por supuesto esta no es la única forma de medir la relevancia ya que no hay un método específico, otros ejemplos son:

- Coeficiente de Jaccard en modelos de indexación binarios.
- Coseno en modelos de indexación vectorial de tipo tf-idf.

## 4. Modelo de arquitectura multi-capa para motores de búsqueda

## 5. Modelo de arquitectura del motor de búsqueda de Google

### 5.1. Características

### 5.2. Metodología

Partiendo del modelo de arquitectura más general, vamos a ver algunos detalles particulares que incluye Google para optimizar los resultados de las búsquedas.

- Las estructuras de datos usadas están optimizadas para manejar grandes colecciones de datos.
- El crawling se realiza usando una gran cantidad de crawlers de forma distribuida.
- Las páginas webs obtenidas mediante el crawling se almacenan comprimidas en un repositorio. La función de indexación descomprime los documentos y los parsea para crear el índice.
- Otro aspecto a tener en cuenta es el texto de los enlaces, ya que normalmente contienen la mejor descripción de la página.

### 5.3. PageRank

Para mejorar la precisión de los resultados obtenidos en la búsqueda, Google hace uso de una estructura de grafo de la red asignando a cada enlace un peso que indica su calidad. De esta forma se calcula el llamado PageRank que proporciona una medida de la importancia de la página en función de las páginas que contienen enlaces a la misma. Este valor se calcula sumando el número de enlaces a la página normalizados por un parámetro  $d$  que indica la importancia de la página de la que procede el enlace.

El PageRank puede interpretarse como el comportamiento de un usuario que navega aleatoriamente. Dicho usuario navega entre las distintas webs a través de los enlaces de dichas páginas. Además se tiene en cuenta que el usuario puede en un momento dado abandonar la página actual. Dada una página  $A$ , un conjunto de páginas  $T_i$  que enlazan a  $A$  y un parámetro de probabilidad de teletransporte  $d$  su PageRank se calcula como:

$$PR(A) = (1 - d) + d \left( \frac{PR(T_1)}{C(T_1)} + \dots + \frac{PR(T_n)}{C(T_n)} \right)$$

## Referencias

1. [http://www.cwr.cl/la-web/2003/stamped/15\\_risvik\\_k-updates.pdf](http://www.cwr.cl/la-web/2003/stamped/15_risvik_k-updates.pdf)
2. <http://www.sciencedirect.com/science/article/pii/S016975529800110X?via%3Dihub>