Ingeniería de sistemas de información
Búsquedas en Intenet

Daniel López García
Lothar Soto Palma
Universidad de Granada

12 de mayo de 2017

Índice

Introducción

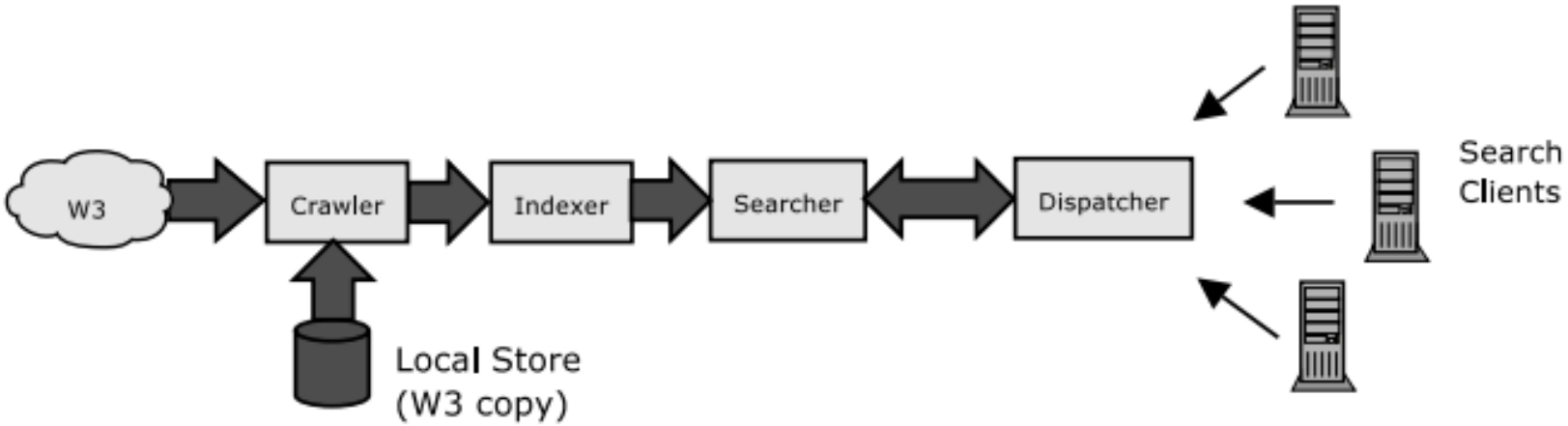
Internet se ha convertido en una fuente de conocimiento masiva, como consecuencia los motores de búsqueda web se han vuelto muy importantes y la necesidad de obtener una arquitectura para estos sistemas que óptimice la realización de búsquedas es ahora una necesidad.

En la web nos encontramos con varios problemas para realizar búsquedas:

- Escalabilidad
- Volatilidad
- Variabilidad

Arquitectura de un buscador.

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

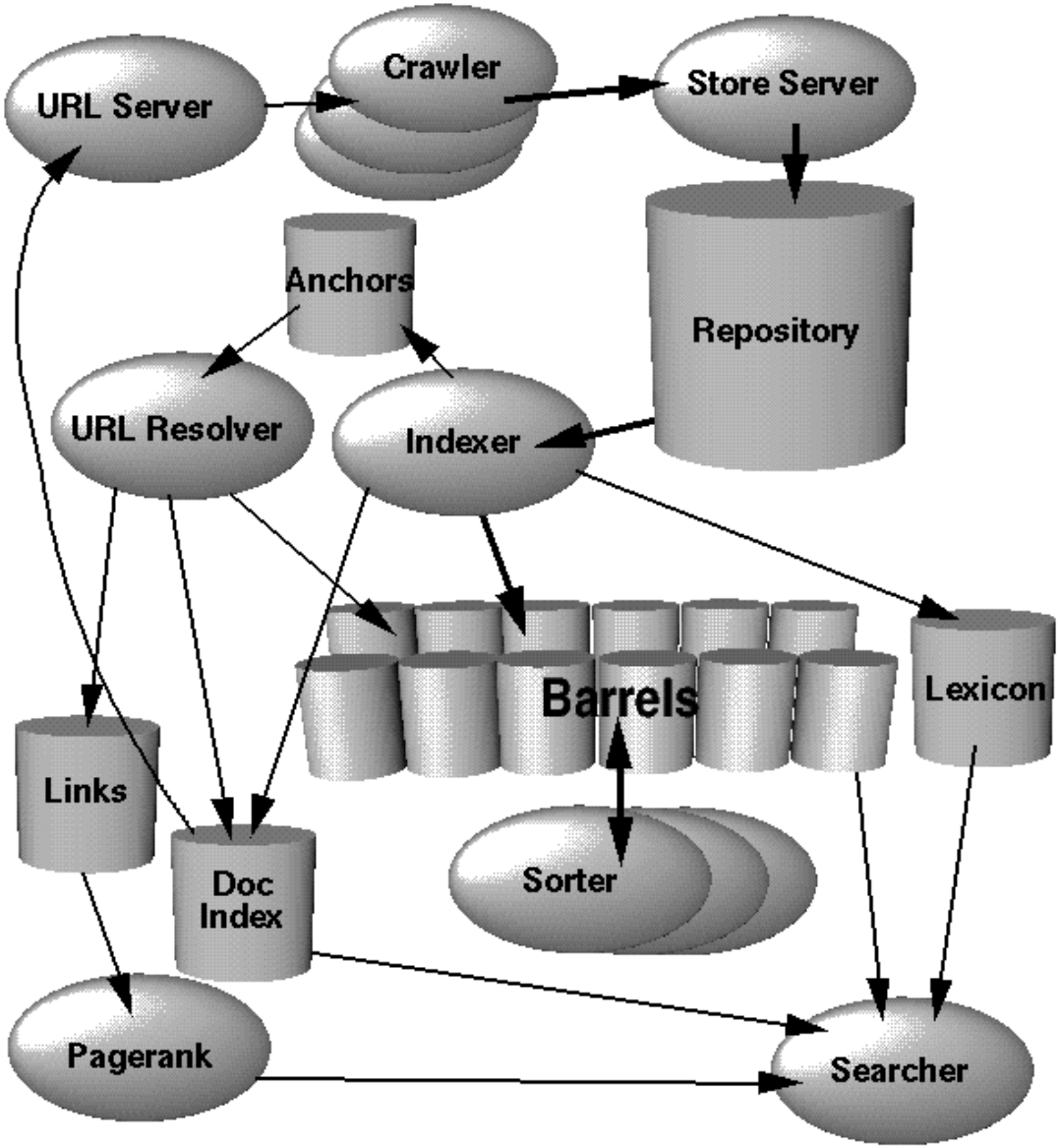


- **Crawler**: Se encarga de la agregación al sistema de los documentos que posteriormente serán indexados para su búsqueda.
- **Indexer**: El indexador es el módulo que se encarga construir una estructura de acceso rápido llamada índice. Existen diversos modelos de indexación algunos de ellos son:
  - Modelo binario.
  - Modelo Vectorial.
  - Modelo Probabilistico.

Arquitectura de un buscador.

- **Searcher**: El buscador es aquel módulo que se encarga realizar las consultas sobre el índice.
- **Dispatcher**: Es el encargado de la recepción de la consulta realizada desde el buscador y enviarla al Searcher para posteriormente obtener una lista ordenada en función de una puntuación denominada relevancia. Algunos métodos para calcular la relevancia son:
  - Coeficiente de Jaccard en modelos de indexación binarios.
  - Coseno en modelos de indexación vectorial de tipo tf-idf.

Arquitectura de Google



- Las estructuras de datos usadas están optimizadas para manejar grandes colecciones de datos.
- Se hace uso del modelo MapReduce para la creación de los índices.
- 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 indice.
- Otro aspecto a tener en cuenta es el texto de los enlaces, ya que normalmente contienen la mejor descripción de la página.

PageRank

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 pagina normalizados por un parámetro d que indica la importancia de la página de la que procede el enlace.

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

Enlaces de referencia.


http://www.cwr.cl/la-web/2003/stamped/15_risvik_k-updates.pdf


http://www.sciencedirect.com/science/article/pii/S016975529800110X?via%3Dihub