Búsquedas en Internet

Ingeniería de sistemas de información

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

Índice

Índice de figuras

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 intenter 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 óptimice 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.

2. Modelo de arquitectura referente de motores de búsqueda

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

- **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 incluyen las URLs anteriores
  - Para cada URL de la cola:
  - Descargamos el contenido de la web
  - Estraemos 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 mayoria 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 en la elaboración de un índice centrandose en la aparición o no de los términos en los documentos.
  - Modelo Vectorial: Este modelo a diferencia del alterior 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 anteriorer llamado modelo tf-idf.
  - Modelo Probabilistico: El modelo probabilistico se centra en la probabilidad de aparición de un término en un documento.

- **Searcher**: El buscador es aquel módulo que normalmente se ejecuta sobre 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 servidor que realiza el cotejamiento con el índice para posteriormente obtener una lista ordenada por una puntuación denominada relevancia de URLs.

3. Modelo de arquitectura del motor de búsqueda de Google

3.1. Características

3.2. 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 indice.
- Otra aspecto a tener en cuenta es el texto de los enlaces, ya que normalmente contienen la mejor descripción de la página.

3.3. 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 pagina 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:

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

Referencias

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

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