Búsquedas en Internet

Ingeniería de sistemas de información

DANIEL LÓPEZ GARCÍA
LOTHAR SOTO PALMA
Universidad de Granada
12 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 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 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.

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. Búsquedas en Internet

Un buscador web tiene que tener conocimiento de todas las webs para satisfacer las consultas de los usuarios la manera de hacer esto puede diferir dependiendo del sistema utilizado, el proceso más común es la construcción de un índice invertido donde las palabras de cada página son añadidas al índice que están enlazadas con los documentos en los que se encontró la palabra. El motor de búsqueda unicamente se encarga de escanear el índice para determinar los documentos que contienen las palabras de las que se compone la consulta y estos se añaden al conjunto de resultados.

Para realizar una consulta sobre Internet es necesario indexar las páginas de las que se compone para ello es necesaria la paralelización y el uso de miles de máquinas, si tuviéramos 5000 máquinas podríamos indexar un billón de páginas aproximadamente en un día pero Internet es varias magnitudes más grande por lo que se necesita aumentar el número de máquinas y una forma inteligente de elegir que páginas webs indexar y cómo.

Pero hay otros problemas como el almacenamiento del índice, normalmente este no puede ser alojado en una única máquina así que la cantidad de datos tiene que ser distribuido entre los discos de cada máquina. Además la distribución de los datos de al almacenamiento también ayuda al rendimiento de la consulta, la tarea de escanear las secciones de las webs y combinar las coincidencias en una etapa posterior.

4. Modelo de arquitectura referente de motores de búsqueda

4.1. Elementos del modelo

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

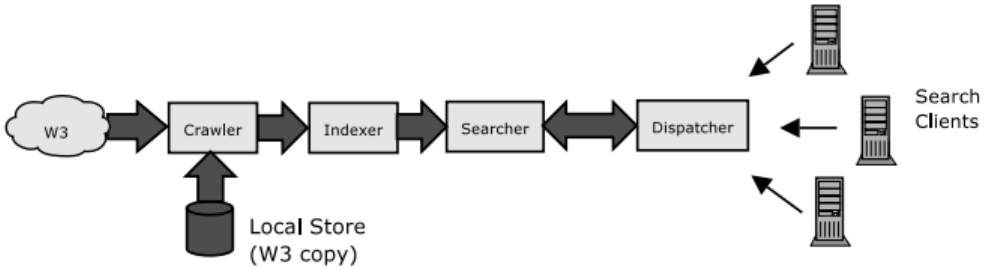

Figura 1: Modelo de referencia

- **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 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 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 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). Otro modelo que combina las dos ideas anteriores es el 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 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.

4.2. Medida de relevancia

El grado de relevancia de un documento web es un 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.

5. Modelo de arquitectura del motor de búsqueda de Google

5.1. 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.
- 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.

5.2. 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:

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

6. Modelo de arquitectura multi-nivel para motores de búsqueda

La principal diferencia de esta arquitectura es la introducción de nodos de búsqueda, un nodo de búsqueda contiene una partición del índice de búsqueda, una parte de todos los datos. Esto produce cambios en la escalabilidad del sistema y en la estructura del mismo.

6.1. Escalabilidad

Dada una agrupación de nodos de búsqueda (cluster) el particionamiento y la replicación son usados para crear una escalabilidad lineal en el tamaño y la valoración de la consulta.

- Particionamiento: El nodo de búsqueda tiene una partición del índice completo. Cuando se recibe una consulta el dispatcher lo que va a hacer es enviar la misma a todos los nodos de búsqueda donde cada uno procesará la consulta y por último se combinarán los resultados de su agrupación. El tiempo de consulta en cada nodo ahora es reducido pasando de tener que analizar todo el índice en el modelo básico a analizar la parte correspondiente a cada nodo de búsqueda, pero el cuello de botella se suele en centrar en el mecanismo de combinación de estos resultados.
- Replicación: A través de la replicación de los nodos de búsqueda podemos incrementar la velocidad de procesamiento de una consulta, el dispatcher ahora usa un algoritmo de asignación round-robin para determinar a que nodo de entre los replicados le enviará la consulta.

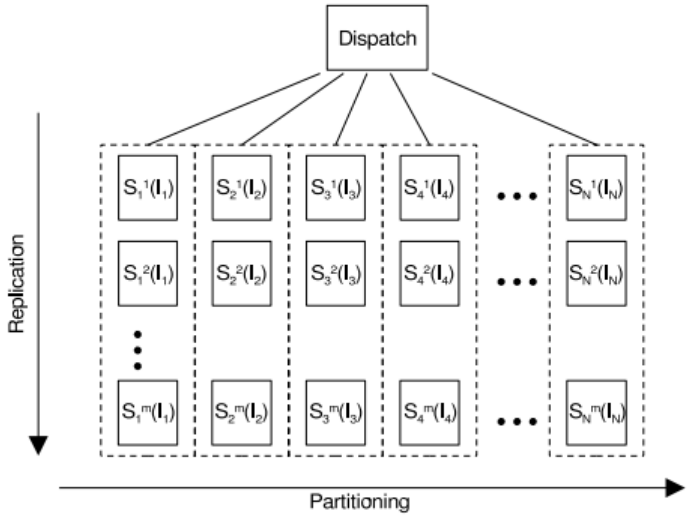

Figura 2: Escalabilidad del modelo

6.2. Arquitectura multinivel

La idea consisten en agrupar los documentos en niveles, cada consulta comienza consultado en el nivel 1 y un algoritmo (fallthrough) decidirá si la consulta continuara ejecutándose en los niveles consecutivos.

6.2.1. Elementos del modelo

- Mapeado de nivel: Dado un conjunto de documentos, existe una función llamada función de mapeado que toma el vector de características de cada documento y lo clasifica en un nivel.
- Alforitmo Fallthrough: Cada consulta comienza en el nivel 1, este algoritmo va a determinar el camino de ejecuciones que seguirá la consulta en los consecutivos niveles en base a su valor de relevancia y número de coincidencias con el índice. Cuando el algoritmo decide que hay que consultar un nuevo nivel entonces el calculo de la relevancia se realiza mediante la combinación de los resultados anteriores.

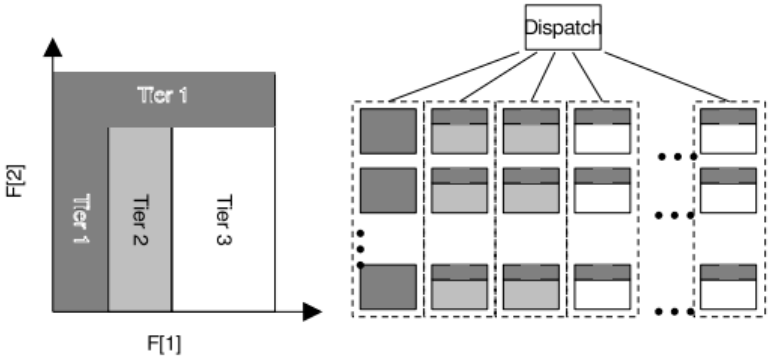

Figura 3: Modelo multi-nivel

6.3. ¿Cómo mejora esta arquitectura a la básica?

El objetivo de esta arquitectura es mejorar el rendimiento del proceso de consultas al sistema, el algoritmo de fallthrough es el encargado de reducir el número de nodos de búsquedas a utilizar (se supone de forma eficiente) y esto lleva a que las consultas no tienen que cotejarse con los índices completos sino con particiones de los mismos. La separación de niveles es aquello que no parece ayudar mucho sin embargo lo que hacen es particionar los contenidos de un documento para descartar aquellos que no tienen relación alguna con la consulta más rápido por ejemplo:

- Nivel 1: Títulos y Links de todos los documentos.
- Nivel 2: Cuerpo de una selección de documentos.

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`