UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE ELECTRÓNICA
REDES DE COMPUTADORES I

14 de Julio 2010

Motores de búsqueda

Elías Escobar 2830031-k

Pablo Petrowitsch 2604325-5

Resumen

En este informe echaremos un vistazo al funcionamiento de los motores de búsqueda en la red mundial de la internet: sistemas informáticos que permiten la recopilación de archivos o páginas web, aumentando la facilidad y accesibilidad con la que se puede obtener la información deseada. Describiremos brevemente su evolución y hacia donde progresa esta, explicaremos las distintas etapas de los procesos relacionados con su operación, los desafíos y dificultades que se deben enfrentar para llevarlos a cabo y hablaremos sobre algunos de los criterios que utilizan para evaluar la relevancia de los documentos.

Introducción

En momentos, en que la internet es la base mundial de Información, lugar de Trabajo y herramienta fundamental para las comunicaciones y el comercio, en donde el subir información a la web nunca fue tan fácil, es necesario el recopilar la información, el almacenarla, y distribuirla de acuerdo a las necesidades e intereses de los usuarios de internet. Por este motivo se crearon los motores de búsqueda, cuya tecnología actual permite obtener en una fracción de segundo miles de resultados sobre casi cualquier tema del que se desee indagar. La mayoría de los usuarios, especialmente aquellos nacidos durante la década de los 90, utilizan estas tecnologías sin imaginar la secuencia de pasos que hay detrás, ni las dificultades que enfrentaron, y siguen enfrentando, los diseñadores de estos sistemas informáticos, sin los cuales la mayor parte de la información contenida en la web sería inútil para el usuario promedio.

En este informe expondremos los mecanismos básicos que le permiten a los motores de búsqueda entregarnos la información que necesitamos, también comentaremos algunos de estos como :

- Google.com
- Alexa Internet
- Bing.com
- Yahoo Search
- Ask.com
- Wolframalpha.com
- Baidu
- Youtube (para multimedia).

Situación Actual

El enorme tamaño de la web y la compleja variedad de necesidades de los usuarios hacen difícil la tarea de encontrar resultados “precisos” a las consultas.

Andrei Border establece 3 etapas de la evolución de los motores de búsqueda:

1 Generación: Solo se considera la información contenida en los documentos como el texto y la estructura.

2º Generación: Los motores de búsqueda ,junto con la información considerada de sus predecesores, estos usan información de análisis de links, anchor-text y del registro histórico de selecciones de los usuarios(“Clickthrough Data ”).

3º Generación : El propósito es aun mas ambicioso, que consiste en responder directamente a la necesidad de las consultas, es decir, entregar resultados de una forma totalmente sensible al tipo de necesidad plasmada en ella. Por ejemplo si se coloca el nombre de una ciudad, el motor de búsqueda debe ser capaz de presentar mapas, sitios de reserva de pasajes de avión , información climática, etc… Es esta generación la que está emergiendo hoy en día.

Funcionamiento

Podemos dividir la operación de los motores de búsqueda en 3 etapas: **Web Crawling**, **Indexado** y la **búsqueda** en sí. Los motores de búsqueda funcionan recopilando información sobre las páginas web, que adquieren de su código HTML. Esta información es obtenida mediante el uso de programas robot llamados arañas, los que, cada cierto tiempo, recorren la web abriendo las páginas como lo haría un browser, descargando todo o parte de su código, comprimiéndolo y siguiendo cada uno de sus links hacia otras páginas, repitiendo el proceso. Posteriormente, se indexan y se clasifican los contenidos de las páginas para determinar como deben ser organizadas. A cada página se le asigna un número que la identifica. Luego, estas se dividen en sus componentes léxicos: las palabras (tokens), se indexan (indexado directo) para saber que palabras aparecen en cada página y guardar algunos datos adicionales sobre los caracteres que allí aparecen. También se le asigna un número de identificación a cada palabra. Finalmente, se reordenan para saber qué páginas o documentos contienen cada palabra (indexado indirecto). A partir de este índice invertido, se obtienen las páginas que se aparecerán cuando se entreguen los resultados de la búsqueda.

Cuando el usuario ingresa el texto con la información que desea buscar (query), el buscador divide el texto en palabras y les asigna a cada palabra el número de identificación que le corresponde, posteriormente, va recorriendo el índice inverso para ver que documentos contienen cada palabra del texto ingresado, hasta que encuentre alguna que los contenga todos. Evalúa la relevancia del documento respecto a la de los otros documentos encontrados anteriormente y entrega una lista con los k documentos más relevantes de la lista.

Dificultades y desafíos implicados:

Por su inmenso volumen, solo una fracción de la red puede ser descargada por las arañas de cada buscador, por lo que se deben priorizar aquellas páginas cuyo contenido pueda ser más relevante para los usuarios. Para hacer esto, cada motor de búsqueda emplea sus propios algoritmos. También se debe considerar que continuamente aparecen y desaparecen páginas de internet, por lo que la información almacenada en el índice puede quedar desactualizada rápidamente. Además, se debe coordinar el funcionamiento simultáneo de varias arañas recorriendo zonas distintas de la red y limitar el ancho de banda que utilizan, ya que cada página envía múltiples peticiones de páginas por segundo y es probable que coincidan múltiples arañas descargando contenidos de un mismo servidor. Para paliar en algo este problema, se ha implementado un protocolo de exclusión de robots, también conocido como robots.txt. Este le indica a la araña que partes del servidor pueden o no ser accedidas por las arañas. Actualmente, las arañas de numerosos buscadores incluyen una política de demoras que establece un tiempo de espera entre peticiones consecutivas de páginas, con el fin de limitar el uso del ancho de banda y no saturar el servidor.

En cuanto al índice, la creación de este debe considerar varios factores como la integración de nuevos datos: saber si está agregando nuevos datos, o si simplemente está actualizando los que ya tiene y mantener la disponibilidad de la información, al tiempo que la actualiza. También se deben comprimir los datos y organizar la información de forma de poder encontrar la información en el menor tiempo posible, utilizar la menor cantidad de recursos, poder hacerle mantención al servidor.

Para que la búsqueda pueda llevarse a cabo, deben poder identificarse las palabras que se desea encontrar, para esto se debe separar los documentos en las palabras que las componen lo que requiere poder identificar el tipo de archivo, el tipo de caracteres y el idioma en el que está escrito el documento e incluso, e algunos motores de búsqueda, la categoría léxica a la que pertenece la palabra.

Rankeo de páginas Web:

Existen distintos métodos para evaluar la relevancia de los documentos en internet: El primero fue, Hyper Search, de Massimo Marchiori, otros algoritmos usados son el HITS (Hyperlink-Induced Topic Search), desarrollado por John Kleinberg y el Pagerank, desarrollado por Larry Page y Sergei Brin. A modo de ejemplo explicaremos el funcionamiento de Pagerank, el algoritmo utilizado por Google: Este algoritmo intenta simular el comportamiento de un usuario que va clickeando links aleatoriamente hasta que se aburre y deja de hacerlo en una página. Esto se puede modelar mediante una cadena de Markov, en la que cada estado representa una página y la probabilidad de que el usuario termine en ella viene dada por el número de enlaces que la apuntan y por el pagerank de las páginas que la enlazan.

El Pagerank, al igual que otros algoritmos basados en el número de enlaces, es un algoritmo iterativo: Supongamos un universo de sólo 4 páginas: A, B, C y D, donde la probabilidad inicial de estar en cada página es la misma y que las páginas B, C y D enlazan a la página A. Sea Pr(A) la probabilidad de llegar a la página A (Pagerank) y L(A) el número de links que salen de la página A, si consideramos sólo un enlace por página tenemos:

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}.$$

Entonces dada una página u, si B es el conjunto de páginas que enlazan a u y v es un elemento de B, el pagerank de u puede ser calculado como:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

Para hacer más realista este modelo, se incluye un factor de damping, este representa la probabildad de que, tras cada cambio de página, el usuario se canse de hacer click y se quede ahí. El factor de damping es sustraído a 1 (y en algunas variaciones del algoritmo, el resultado es dividido por el número de documentos N) y este término se agrega luego al producto del factor de damping y la suma de los puntajes de PageRank que recibe la página de aquellas que la enlazan.

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right).$$

Por lo tanto, el Pagerank de cualquier página es derivado en gran parte de los Pageranks de otras páginas. El factor de damping hace tender este valor hacia abajo.

Motores de búsqueda más famosos:

- **Google: Tiene** el buscador más exhaustivo del mundo y presta servicios de búsqueda en la Web a más de 100 países. El algoritmo del ranking de búsquedas es el siguiente :

GoogScore = (Uso de palabras clave en los contenidos * 0.3) + (Relevancia del dominio * 0.25) + (Calidad de los links entrantes * 0.25) + (Comportamiento de los usuarios * 0.1) + (Calidad del contenido * 0.1) + (Puntuación manual) – (Penalidades)

- **Yahoo! :** En febrero de 1994 diseñaron un sitio en Internet que catalogaba otros sitios de interés, a fin de tener acceso periódico a ellos de forma rápida y sencilla. Almacenaban las listas en categorías y estás, en su categorías, este es el concepto básico de yahoo! .Otros servicios: Correo electrónico, Salas de charla, mapas y rutas de carretera, Filtros contra correo basura y programas antivirus.

- **Baidu** (百度 en chino): Es un motor de búsqueda chino, Su diseño es similar al de Google e incluye la posibilidad de búsqueda de noticias, imágenes y canciones, entre otras funciones. Es el 6to sitio más visitado de Internet

- **Youtube:** Para almacenar y distribuir la información, han construido un tipo de FAT (Tabla de Asignación de Archivos) en los discos duros. YouTube segmenta el espacio de forma más pequeña de lo habitual. Esto ejemplo permite ahorrar mas espacio en disco, permite optimizar los medios de almacenamiento de forma más eficaz, y permite la lectura más aprisa de los datos.

Conclusiones

Los actuales motores de búsqueda nos permiten acceder fácilmente a casi cualquier contenido de la red. Esto es posible gracias al empleo de programas araña, que recorren la web recolectando y organizando la información disponible sobre ellas, y a la creación de los índices, que permiten referenciar las páginas a partir de las palabras que contienen. Pero esto no sería posible sin el desarrollo de las ténicas de reconocimiento de palabras, de protocolos adecuados para reconocer formatos de archivos o documentos y de los algoritmos de rankeo de las páginas web. Se ha dicho que internet ha provocado la mayor revolución cultural desde la invención de la imprenta, pues no sólo permite disponer de una cantidad ilimitada de información en todo momento, si no que también permite acceder a ella de forma instántanea. Si bien no todos concuerdan en que esto resulte ser bueno a largo plazo, no cabe duda de nada de esto sería posible sin el desarrollo de los motores de búsqueda.

Anexos

Bibliografía

[1] Juan Francisco Zamora Osorio (UTFSM) *"Construcción de algoritmos de identificación automática de necesidades de usuarios en motores de búsqueda"* Tesis magister en ciencias de la ingeniería informática .

[2] http://86400.es/2006/10/31/pero-como-funciona-google/

[3] http://es.wikipedia.org/wiki/Google

[4] http://wappy.ws/softwarecomo-funciona-youtube-20070809.html

[5] Sergey Brin y Lawrence Page: "The Anatomy of a Large-Scale Hypertextual Web Search Engine" (sección 4 "System Anatomy").

[6]http://en.wikipedia.org/wiki/Pagerank

[7] http://en.wikipedia.org/wiki/Search_engine

[8]http://en.wikipedia.org/wiki/Web_crawling

[9]http://en.wikipedia.org/wiki/Index_(search_engine)

[10]http://en.wikipedia.org/wiki/Web_search_query