

Búsquedas en Internet

Ingeniería de sistemas de información

DANIEL LÓPEZ GARCÍA
LOTHAR SOTO PALMA
Universidad de Granada
28 de abril de 2017

Índice

1. Introducción	2
2. Modelo de arquitectura referente de motores de búsqueda	2
2.1. Características del modelo	2
2.2. Elementos del modelo	2
2.3. Metodología	2
3. Modelo de arquitectura del motor de búsqueda de Google	2
3.1. Características	2
3.2. Metodología	2

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 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.

2. Modelo de arquitectura referente de motores de búsqueda

2.1. Características del modelo

2.2. Elementos del modelo

2.3. Metodología

3. Modelo de arquitectura del motor de búsqueda de Google

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.

3.1. Características

3.2. Metodología

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

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.

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>