CS101 Lecture 29:

Web Traffic - pct of Page Views

Source: alexa.com, 4/5/2010
What You’ll Learn Today

– What is a search engine?
– How does Google enable web searches?
– What processes are necessary to make Google work?

Search Engine Basics

A search engine finds web pages for clients.

The client enters a query (search terms), and the search engine returns a list of pages relevant to the query.
Google’s Building Blocks

Google’s search engine is based on 3 major components/processes:
– Web Crawler
– Indexer
– Query Processor

Googlebot

A web crawler is an automated process (a “bot”) which downloads web pages, following links from one page to the next.

Google’s web crawler is called Googlebot.
– Finds pages on the world wide web, and hands them off to the Indexer.
Googlebot

Googlebot consists of many computers capable of fetching thousands of pages simultaneously.

Two ways Googlebot learns about pages:
- Add Your URL
- Following links from known pages

www.google.com/addurl.html
Issues in Webcrawling

Some challenging issues in web crawling:
- Don’t revisit the same page unnecessarily! This wastes resources.
- How often to visit a page to check for updates?

Google Indexer

Googlebot gives the Indexer the full text and HTML of every page it visits.
- Pages indexed by search terms. Which words?
- The Indexer ignores some common words. Which ones? Why?

- What is the process of creating an index?
- How much data gets indexed?
- How long will it take to create that index?
Google’s Query Processor

The Google Query Processor has three parts:
User interface
Evaluation engine
Results formatter

Query Processor: Advanced Search

Find web pages that have...
all these words:
this exact wording or phrase:
one or more of these words:
But don’t show pages that have...
any of these unwanted words:

Need more tools?
Results per page: 10 results
Language: any language
File type: any format
Search within a site or domain: (e.g. youtube.com, .edu)

* Date, usage rights, numeric range, and more
Query Processor: Evaluation

The evaluation engine searches the indexed database for the search terms and develops a PageRank.

Google considers over 100 factors to calculate PageRank.
- Popularity of the page
- Position of search terms in page

The Google Query Process

1. The web server sends the query to the index servers. The content inside the index servers is similar to the index in the back of a book—it tells which pages contain the words that match any particular query term.

2. The query travels to the doc servers, which actually retrieve the stored documents. Snippets are generated to describe each search result.

3. The search results are returned to the user in a fraction of a second.

Copyright © 2003 Google Inc. Used with permission.
What You Learned Today

– Web Crawling
– Indexing
– Query Processing
– Remaining Questions
  • How does query processor search the index?
  • How long does it take to search?
  • What is the process of creating an index?
  • How long will it take to create that index?

Announcements and To Do List

– HW12, due 4/20
– Readings:
  • http://www.googleguide.com/google_works.html (today)
  • http://www.sorting-algorithms.com/ (Friday)