NUST

NATIONAL UNIVERSITY
OF SCIENCES & TECHNOLOGY

BESE-13

Assignment 2
Summary of "The Anatomy of a Large Scale Search Engine"

Student: Sharjeel Sajid
CMS: 407285
School Of Electrical Engineering and Computer Science
National University of Science and Technology

Contents

Title:

The Anatomy of Large-Scale HyperTextual Web Search Engine

Authors:

Sergey Brin and Lawrence Page

Summary

1. **Introduction**

   This research paper introduces Google as a large-scale search engine designed to address the challenges presented by the rapid growth of the web, emphasizing the importance of precision in search results. It highlights the need for high-quality search outcomes, as merely having a complete index does not guarantee user satisfaction. Additionally, the paper underscores Google's commitment to fostering academic research and data accessibility, enabling innovative experiments and studies on large-scale web data. Google's design aims to scale with the web's growth by employing efficient crawling technology, storage management, and fast query handling while leveraging the structure of hypertext to enhance search quality and relevance.

2. **Features**
   There are 2 very important features in Google

   (a) **Page Rank**
   Google has a method for ranking pages. By ranking pages, Google is able to display the important results at the top. Page Rank views web links as citations. When one-page links to another, it's essentially endorsing or citing the linked page. The more citations (links) a page receives, the more authoritative and important it's considered. PageRank is calculated iteratively and corresponds to the principal eigenvector of the normalized link matrix of the web.

   (b) **Anchors** Google uses anchor text associated with links, not only with the source page but also with the destination page, providing more accurate page descriptions and allowing retrieval of uncrawlable content like images or databases. Propagating anchor text to target pages expands search coverage and enhances result quality.

   These features, in combination with proximity and visual presentation considerations, contribute to Google's capability to deliver highly relevant and efficient search results on the web.

3. **System Anatomy**

   (a) **Google Architecture Overview** The process to organize and search the web in Google is explained below.
   There are "crawlers" that visit web pages and send them to a storage server. Each page gets a unique ID(DocID). Then, the "indexer" reads the pages, looks at the words, and keeps track of where they are in the document. It also stores information about the links on each page. A special tool called the "URL resolver" converts web

addresses into these unique IDs. This way, Google can find where the links go. The "sorter" organizes all this information so it's easy to search. The web server uses this sorted data, along with Page Ranks to answer the search queries.

(b) **Data Structures** The major data structures used in Google to crawl, index and search data quickly are explained below.

i. **Big Files** In Google Big Files are stored in multiple files and are given a 64 bit integer address. This allocation of files in multiple file system is handled automatically. Big Files packages also handles alloation of File descriptors and compression operations.

ii. **Repository** In repository full HTML document is stored along with DocID, Length and URL. The HTML document is compressed using zlib before storing.

iii. **Document Index** Document Index stores information of each document. It stores document status, a pointer to document stored in repository, a document checksum. It also contains a pointer to a file containing URL and title if it has been crawled. The URL can be converted to DocID by calculating the URL's checksum and then comparing it in a file containing all the checksums and DocIDs.

iv. **Lexicon** Lexicon is a collection of commonly used words. This collection helps to perform search operations. Google lexicon has 14 million words. Lexicon contain words along with a hash tables of pointers.For every valid wordID, the lexicon contains a pointer into the barrel that wordID falls into.

v. **Hit Lists** Hit lists store the occurrence of a certain word in a document. There are two types of Hit lists Fancy and Plain. Fancy hits represent anchors, titles, or meta tags. The hit list also contains the font size, capitalization, and position which are encoded using a compact encoding. It consists of 16 bits. The first bit represents capitalization, the next 3 bits represent font size and there is a 12 bit for position in the document. If font bits are 111 then it is a fancy bit. The position bits of fancy bits are as 4 bits for a position in the anchor and 8 bits for a position in the document.

vi. **Forward Index** The forward index contains a range of WordIDs. It also stores the DocIDs of the documents in which a word occures and a hitlists of the word in that document. These are stored in a number of barrels.

vii. **Inverted Index** Inverted index contains the barrels of forward index after they are sorted by sorter. Google has two sets of barrels. In one set all the anchor and tiles are stored sorted by their rankings and the other set contain all the hit lists.

(c) **Crawling The Web** Crawling the web means going from one web page to another and gathering information about web pages. It is one of the most complex task because there are many different types of web pages and many different types of errors occurs which must be handled. One major performance problem is DNS lookup. Google uses distributed crawler system for web crawling to increase performance

(d) **Indexing** The indexing of documents occurs in following steps. First the document is parsed then all the words are stored in barrels and given a wordID. After this they are sorted and converted to inverted index.

4. **Searching** Searching takes place in following steps. Firstly the query is parsed and words are converted to wordIDs. Then the short barrel is searched for matching query and then long barrels. The documents that match the search are displayed according to their page ranks.

5. **Results and Performance** The primary measure of a search engine's quality is the relevance of its search results, and Google has demonstrated its prowess by outperforming major commercial search engines for most queries. For instance, when searching for "bill clinton," Google's results are well-clustered by the server, making it easier to navigate through the outcomes. Google relies on the proximity of words in documents to ensure the relevance of results. Storage efficiency is another focus, with the total data storage size being approximately 55 GB. Notably, the system's performance in crawling, indexing, and sorting is robust. It took around 9 days to download 26 million pages, with the final 11 million pages being fetched in just 63 hours. Google's search performance is decent, answering most queries in 1 to 10 seconds, mainly due to disk I/O. The future objective is to enhance Google's speed and performance, aiming to handle several hundred queries per second.