

Understanding Google's PageRank Algorithm

Salveen Singh Dutt

January 21, 2025

Overview of Google Search Engine

- ▶ Google is a scalable search engine for the World Wide Web.
- ▶ It uses hyperlink structure and content analysis to improve search quality.
- ▶ Core innovations include:
 - ▶ PageRank Algorithm
 - ▶ Anchor Text Usage
 - ▶ Efficient Indexing and Crawling Systems

PageRank Algorithm: Core Idea

- ▶ PageRank evaluates the importance of web pages using their link structure.
- ▶ A page's rank depends on:
 - ▶ The number and quality of links pointing to it.
 - ▶ The ranks of linking pages.

$$PR(A) = (1 - d) + d \sum_{i=1}^n \frac{PR(T_i)}{C(T_i)}$$

- ▶ $PR(A)$: PageRank of page A .
- ▶ d : Damping factor (e.g., 0.85).
- ▶ T_i : Pages linking to A .
- ▶ $C(T_i)$: Outbound links count of T_i .

PageRank Algorithm: Intuition

- ▶ Models a "random surfer" who clicks links at random:
 - ▶ Probability d : Continues surfing.
 - ▶ Probability $1 - d$: Jumps to a random page.
- ▶ inbound links and links from high-ranked pages).

System Architecture

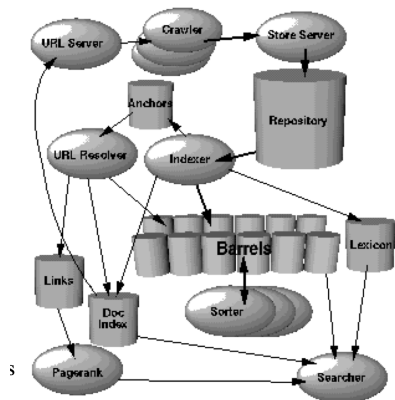


Figure: High Level Google Architecture. (Source: The Anatomy of a Large-Scale Hypertextual Web Search Engine)

System Architecture

- ▶ Components:
 - ▶ Distributed Crawlers: Collect web pages.
 - ▶ Indexer: Parses pages into words and links.
 - ▶ Sorter: Creates an inverted index for fast queries.
- ▶ Efficient Data Structures:
 - ▶ Forward Index: Organizes words by document.
 - ▶ Inverted Index: Organizes documents by word.
 - ▶ Anchor Text: Enhances relevance with descriptive links.

Google Architecture Overview

Key Components and Workflow:

▶ **Distributed Crawlers:**

- ▶ Download web pages from lists provided by the URLserver.

▶ **Storeserver:**

- ▶ Compresses and stores fetched web pages in a repository.
- ▶ Assigns a unique identifier, *docID*, to each page.

▶ **Indexer:**

- ▶ Reads and uncompresses repository data.
- ▶ Parses documents into word occurrences (*hits*) with metadata (position, font size, capitalization).
- ▶ Distributes hits into *barrels* (partially sorted forward index).
- ▶ Extracts links from web pages, saving link structure and anchor text into an *anchors file*.

Google Architecture Overview

URLresolver:

- ▶ Converts relative URLs into absolute URLs and assigns *docIDs*.
- ▶ Associates anchor text with target *docIDs* in the forward index.
- ▶ Generates a links database used for computing PageRank.

Sorter:

- ▶ Resorts barrels by *wordID* to produce the inverted index.
- ▶ Works in place to minimize temporary space usage.

Searcher:

- ▶ Uses the inverted index, PageRank, and a lexicon generated by DumpLexicon to handle search queries efficiently.

Anchor Text and Its Role in Google's Algorithm

What is Anchor Text?

- ▶ The visible, clickable text of a hyperlink.
- ▶ Describes the target page's content.

How Does the Algorithm Use Anchor Text?

- ▶ *Relevance Association:*
 - ▶ Anchor text is associated with the page it links to, improving the page's ranking for relevant terms.
- ▶ *Improved Description for Non-Text Content:*
 - ▶ Links to non-indexable content (e.g., images, videos) provide valuable metadata.
- ▶ *Spam Resistance:*
 - ▶ Difficult for spammers to manipulate anchor text across diverse, reputable sites.

Key Features of Google

- ▶ High Precision:
 - ▶ Combines PageRank, anchor text, and proximity data.
- ▶ Scalable to billions of documents:
 - ▶ Efficient storage and indexing methods.
 - ▶ Parallelized crawling and sorting.
- ▶ Robust against manipulation:
 - ▶ Link structure resists spamming attempts.

References

Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems*, vol. 30, 1998.