# THE ANATOMY OF LARGE-SCALE HYPERTEXTUAL WEB SEARCH ENGINES

ADARSH

[20MCA1459@cuchd.in](mailto:20MCA1459@CUCHD.IN?subject=Regarding%20the%20anatomy%20of%20large%20scale%20web%20search%20engine)

Department of Computer Science, Chandigarh University

Mohali, PUNJAB -140413

**ABSTRACT**

In this paper, I have tried to present the working of Web Search Engine specifically taking Google as an example, a prototype of a large-scale search engine after going through various studies and researches, and also the paper made by *Larry Page* and *Seige Brin*, the founder of **Google**, who proposed their paper in Stanford University back when they started Google, which makes heavy use of the structure present in hypertext. Google was designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 to hundreds of millions of web pages involving a comparable number of distinct terms. It can answer tens of millions of queries every day.

In this digital era, search engine is the main source of information be it anyone from any field. With the development of Internet, there’s a sea of unlimited information which consists of all types of texts, images, multimedia, and other forms of electronic information asset and resources. This paper elaborately describes that how tasks are performed by the search engines. An overview of the whole working of the Search Engine System. It also addresses the structure of retrieval system and to improve its performance through the new emerging technologies and techniques like SEO which indirectly helps in more efficient crawling and indexing.

This paper along with its abstract can be found online at [GitHub](https://github.com/adar5h/The-Anatomy-of-Large-Scale-Hypertextual-Web-Search-Engines).

**Keywords:** World Wide Web, Web searching, Search engines, Web Crawlers, Indexing, metadata.

**INTRODUCTION**

Web based search engines are special sites on the Web that are designed to help people find information stored on various different sites. The Internet has grown rapidly and even after Google being one of the most used search engines as of now, there are other popular search engines too. There are differences in the ways various search engines work, but they all usually perform in four-step process:

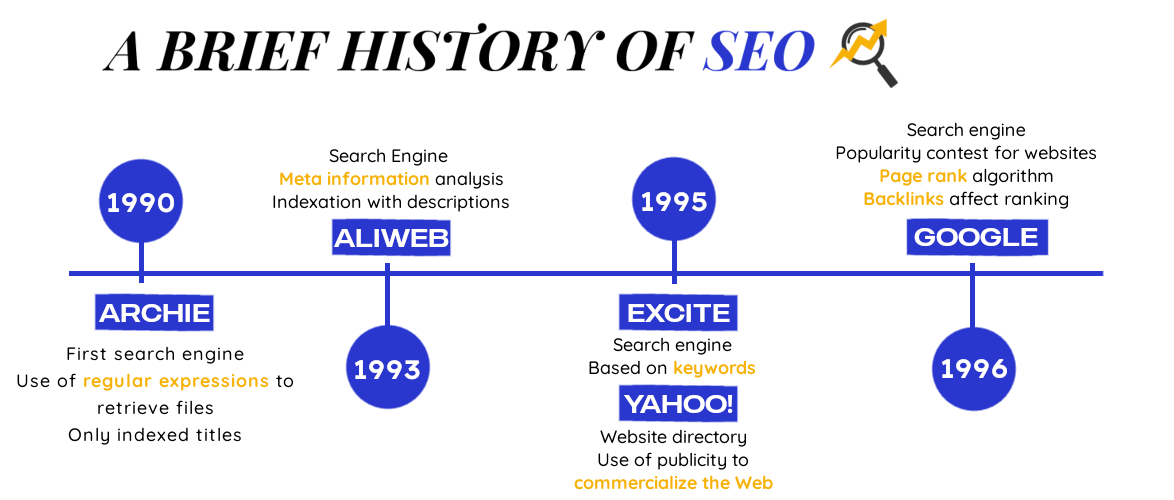
1. Crawling: is the process of finding new pages or updated pages to the search engine.

2. Indexing: Indexing the pages to create an index from every word to every place it occurs.

3. Ranking: Giving rank to the pages so the best ones show up first.

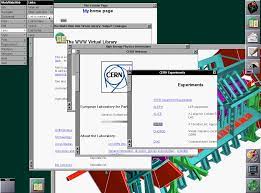
4. Displaying the results in a way that is easy for the user to understand.

During the emergence of *World Wide Web*, the early stage in development of search engines, at time when Yahoo! was on its peak and Google is yet to be established as it is now, search engines held an index of a few hundred thousand pages and documents, and received maybe one or two thousand inquiries each day. Today, a top tier search engine will index hundreds of millions of pages, and respond to tens of millions of queries per day. In 1990, Tim Berners-Lee created the first Web browser (and Web editor) originally called the **World Wide Web** and later renamed to **Nexus** in order to avoid ―confusion between the program and the abstract information space (which is now spelled World Wide Web with spaces); it was written in Objective-C using the NeXT Computer. And at the time, this was the only way to browse the web. You can see a screenshot of the brief history of SEO ,first search engine ,first web browser below.





**First Web Search Engine – Archie.**

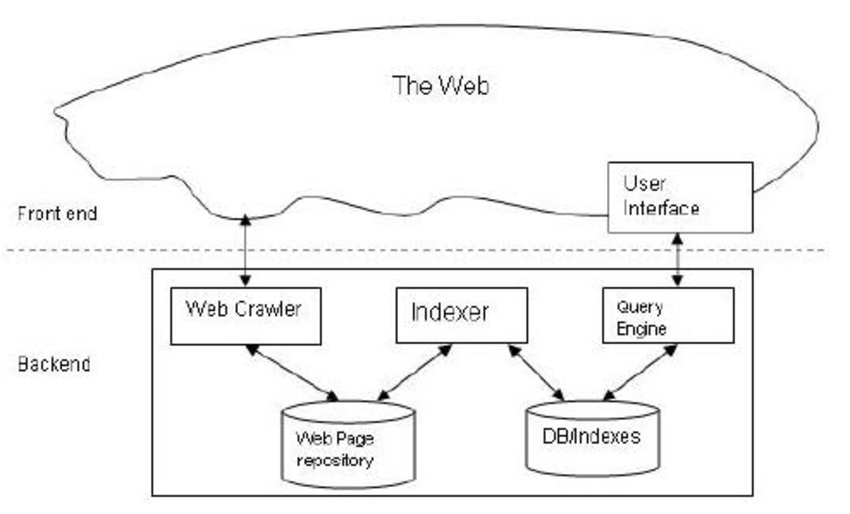


**First Web Browser.**

It was just the beginning in the area of Web Browsers**.** Marc Andreessen, National Centre for Supercomputing Applications (NCSA) at University of Illinois, introduced the Mosaic browser. After this Microsoft also got into the browser business and started rolling out Internet Explorer with their OS. Till date so many search engines have been developed and some of them are active whiles others due to various reasons are now not active.

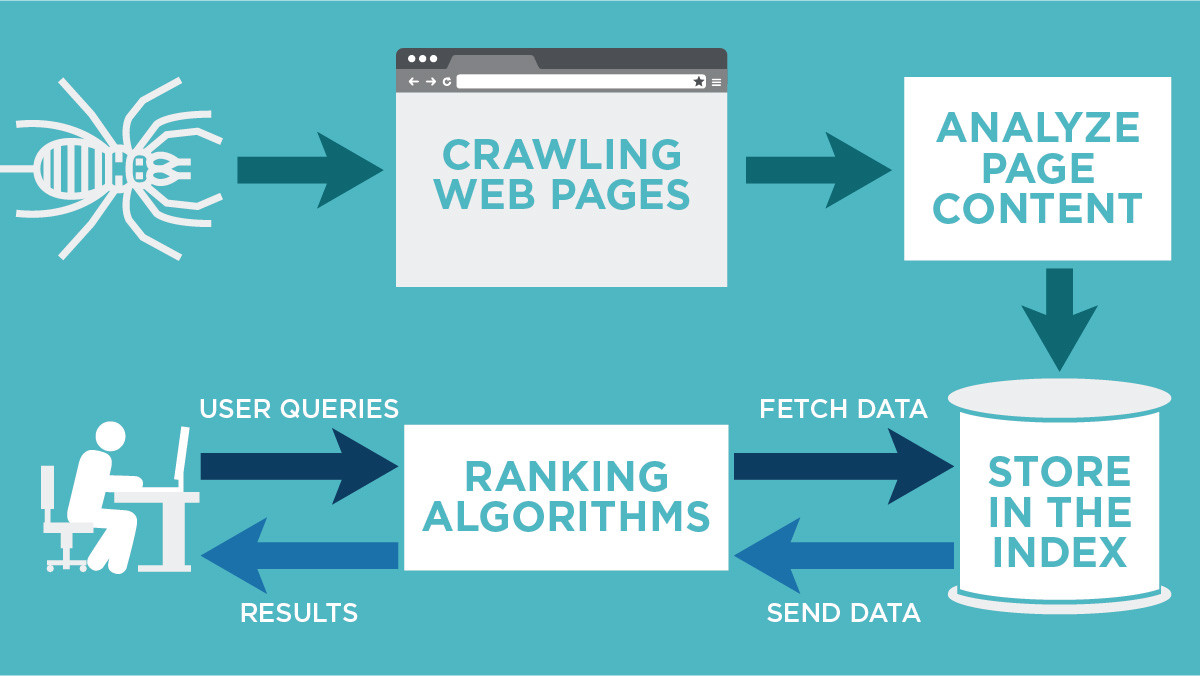
**ARCHITECTURE OF WEB SEARCH ENGINES**

A brief explanation of what Search Engines are programs that searches the web, consisting of web pages, documents, blog or articles for the particular keywords that were typed and searched by the user. The word is pretty much self-explanatory ‘Seach Engine’ – A collection of programs, i.e., an Engine that enable the users to search for their queries on the Internet a.k.a. World Wide Web. The term Search Engine is often used to describe systems like Google, Bing and Yahoo! Search.



**Fig.** Diagrammatic representation of Architecture of Web Search Engine.

The main components and tasks of web Search engines, **Crawling** or **Spidering** could be described as an automated process to gather the data with web spiders. They can be virtualized as little spider and are also known as crawlers and have others name also, known as crawlers, software agents, wanderers, robots, web agents, walkers or know bots, named after their special software robots. This type of search service is called Crawler based or Spider based search engine. Spiders process the web page and give us the information that they have fetched. These web pages are found by them by URL which is given by a web page holder to notify their web page, or through hypertext links embedded that are mostly used in web pages, these days. And in the latter case, spiders start by crawling a few web pages and then from there it follows on the links on those pages. After the process of fetching from the pages they are pointed to, they follow the links that were on the last visited pages. The same process will be continued until the spider have had indexed a certain part of the web that include the pages they store on various machines, which leads to next tasks.



**Fig.** An overview of the crawling process and how it helps the users to get result queries.

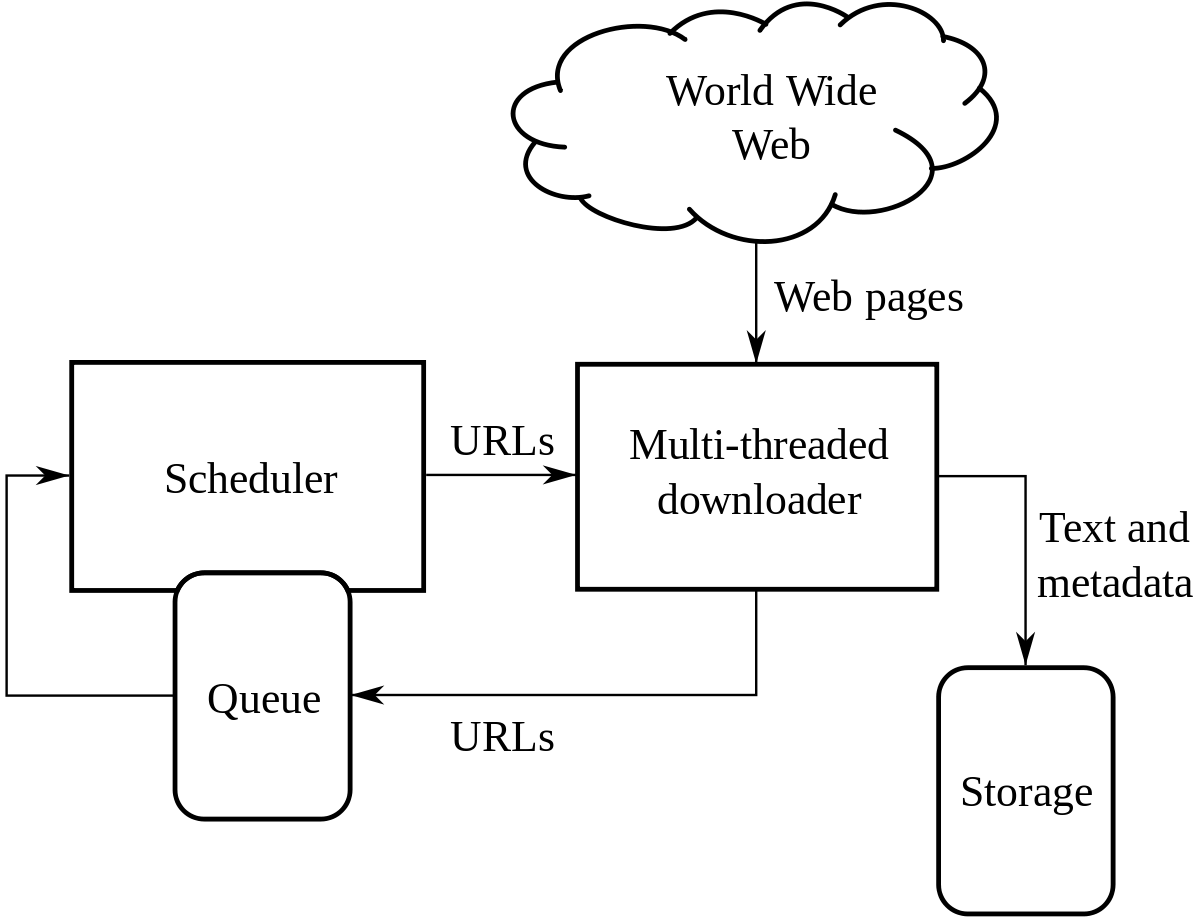
**Indexing** is the second step in the working of web search engines. Basically, it is the process of taking raw data and categorizing it on the basis of various factors, removing duplicate information, thus organizing it all into a systematic accessible structure. Then this stored and full indexed of the crawled web pages are organized in a database, usually in an inverted index data structure. This is the basic architecture of how web-based Search Engine works. It is efficient for *keyword*-*based* queries, in such a way that the documents consisting of the typed keywords can be quickly retrieved. Webmasters have reverse-engineered this whole process and thus have taken the advantage of it to its full potential, mostly for business purposes.

Here, comes the term **SEO** (Search Engine Optimization) or maximizing search engine visibility of web pages, online marketing strategies for making your brand established. We will cover this topic in later part of this paper.

**TOOLS FOR WEB BASED RETREIVAL**

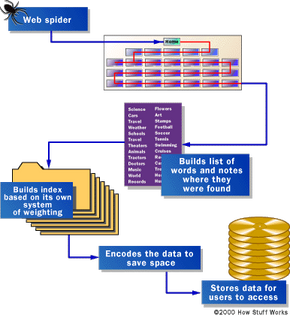
**Web Crawlers/Spiders/Robots**

A web crawler or spider or bots, is an automated program that follow the search engines algorithm to examine the content and structure of pages. Once given a seed or more URLs, it will download the web pages associated, extracts any hyperlinks given in the page and recursively continues to download the web pages identified by these hyperlinks in a systematic manner. A web crawler starts with the given list of URLs (Seeds). The crawler visits these URLs, looks for any hyperlinks in the page and adds them to the list of URLs yet to be visited (crawl frontier). And the same process keeps on happening recursively, URLs from the crawl frontier are visited and then any hyperlinks from them gets added to the frontier and this all happens according to a set of policies set up by the search engine.



**Fig.** Architecture of a web crawler

e.g., Google wants you to include the following information for it to easier understand your content; **title tags, meta descriptions, headings, content, and more**. The reason for this is so that the bots can interpret the content, categories, and products within a page. There are different strategies that can be placed within the code to make sure that the bots are able to crawl a page in the most effectively and efficiently.



**Fig.** Diagrammatic Representation of a web spider crawling in Web. [[Source](file:///C:\Users\iadar\AppData\Roaming\Microsoft\Word\%20https\computer.howstuffworks.com\internet\basics\search-engine1.htm)]

The spider’s initial points are usually the lists of servers with a massive traffic and some very popular pages. The spider begins with indexing the words on the page and following every link found within the site. And within very less time the spidering system quickly begins to travel, spreading out across the most widely used portions of the Web. Google began as an academic search engine initially. In the paper that describes how the system was built, Sergey Brin and L. Page had explained how quickly the spiders built at that time were working. They built their initial system to use multiple spiders, often three at a given time. Each spider could hold up about 300 connections to Web pages open at a time. And during its peak performance, usually using 4 spiders, it could crawl over 120 pages per second, generating 600 kb of data each second. During the early days of Google, they had a server dedicated to providing URLs to spider. Google Spider looks at an HTML page, and would take a note of the things mentioned below:

1. The words within the page
2. Where the words were found

These different approaches usually attempt to make the spider operate faster, allow users to search more efficiently, or both. For example, some spiders will keep track of the words in the title, sub-headings and links, along with the 100 most frequently used words on the page and each word in the first 20 lines of text.

As we are also discussing the ways to improve ranking, indexing, crawling in this paper, here is a guide to help Google per se, to increase the ranking of your website with better crawling and indexing.

* Create a sitemap – holds a complete list of a website’s pages and primarily lets bots know what to crawl
* Add Schema – a “roadmap” for the bots to crawl a page productively
* Disallow content within the Robots.txt file that isn’t necessary to search
* Site speed – if a page loads too slow the robot will leave before it can crawl the full page

They are most commonly referred to as crawlers, but are also known as ants, automatic indexers, bots, spiders, Web robots and worms.

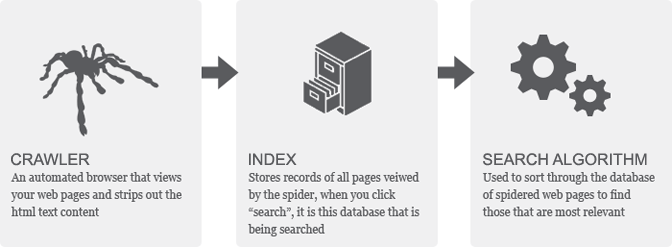
NOTE: *The amount of data that a search engine can store is limited by the amount of data it can retrieve for search results.*

**Indexing**

Once a search engine system is done crawling through a webpage, the next step is to index the crawled page. Indexing is a process in which the search engine stores information about web pages, enabling them to return relevant, high-quality results in some fractions of seconds.

The term Index is usually explained as anything that serves to guide, pin point or to make it easier to associate reference. Although the term Indexing is used here in the same context of ranking and retrieval, it has a precise meaning here. According to the experts – The most important of the tools for information retrieval is the index—a collection of terms with pointers to places where information about documents can be found [Manber 1999]. Indexing could also be described as building a data structure, whose functionality will be beneficial allowing it to search the given text quickly.

Thus concluding, Indexing is the process by which search engines organize information after crawling the webpage, but before a search to enable instantaneous response to queries. Searching through individual pages for keyword and topics is a slow process, considering the massive amount of data World wide web is storing at this point, which in the near future is going to increase only. Therefore, search engines (Including Google) use an **inverted index**, also known as a **reverse index**.



**SEARCHING/SORTING ALGORITHMS**

1. **INVERTED INDEXING**

An Inverted index is a system wherein a database of text elements is compiled along with pointers to the documents which contain those elements. Then, search engines make use of a process called **tokenization** to reduce words to their core meaning, thus reducing the number of resources needed to store and retrieve data. This is a much faster approach that listing all the documents containing the relevant characters and keywords.

### **An example of inverted indexing**

Given below is a basic example which illustrates the concept of inverted indexing. In the example we can see that each keyword (or token) is associated with a row of documents in which that element was identified.

|  |  |  |  |
| --- | --- | --- | --- |
| **Keyword** | **Document Path 1** | **Document Path 2** | **Document Path 3** |
| **SEO** | example.com/seo-tips | moz.com | … |
| **HTTPS** | deepcrawl.co.uk/https-speed | example.com/https-future | … |

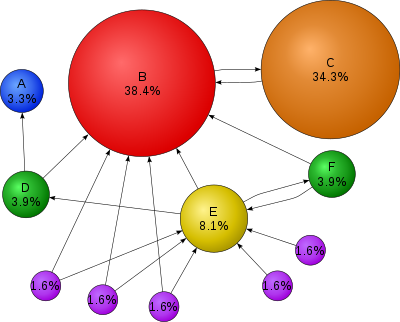
This example uses URLs but these might be document IDs instead depending on how the search engine is structured.

Also, indexing can also be done in other ways. Four approaches to indexing documents on the Web are:

* Human or manual indexing;
* Automatic indexing;
* Intelligent or agent-based indexing;
* Metadata, RDF, and annotation-based.

1. **PAGERANK**

PageRank is a search algorithm, its utility could be known by the fact that Google uses it. It works by counting the number and quality of the hyperlinks in a page to determine the traffic and genuine usage of a website. With the underlying assumption that, the more important a website is, the more likely is it to receive more links from other websites.

****

**Fig.** Graphical Representation of PageRank.

Here circle B is large because of its massive traffic and the linkage to other pages.

The main aim of Search Engine is to give the most convenient and relevant answers to user search queries. Crawling and indexing can help exponentially with Organic search. But in order to make sure that your website has the best SEO practices implemented, it is crucial to ensure that they are being crawled and indexed efficiently & effectively, even the newly added pages and content. This way you can show up within searches that are relative and relatable to user’s intent.

**SEARCH ENGINE OPTIMIZATION**

The SEO is an abbreviation for Search Engine Optimization. As the very name specifies, it is a process of optimizing a website or page to rank higher and come up on the top of the search results.

The key difference between SEO and paid advertising/sponsoring is that SEO involves “organic ranking”, which means you are not paying to be in that space.

Search engines such as [Google](https://searchengineland.com/library/google) and [Bing](https://searchengineland.com/library/bing) use bots to crawl pages on the web, going from site to site, collecting information about those pages and putting them in an index. Think of the index like a giant library where a librarian can pull up a book (or a web page) to help you find exactly what you’re looking for at the time.

Think of the index like a giant library where a librarian can help you search for finding exactly what you’re looking for. Then the algo analyses the page, determining the order in which pages should appear when searched for a given query.

Unlike paid/sponsored ads, we can’t directly get in contact with search engines or pay them to get our ranking, for that motive only organic methods work. The newly update SEO table not just increases the ranking but also has a list of toxins that detract from SEO practices. These are shortcuts or tricks used to work in the early times, but in this age and time the search engines are much more sophisticated. And affects your ranking badly, if you get caught using such shortcuts.

With time SEO is getting divided into niches sections that deep-dives into the SEO: such as Local SEO, Ecommerce SEO. Thus, making it easier to get exactly the kind of organic traffic you want.

**LITERATURE REVIEW**

The main aim of researching on ‘The anatomy of a Large scale hypertextual web search engine’ was to make the understanding of Search Engine Systems clear. By doing so, one can excel and have a strong online presence; it might be to make a brand of yourself as an entrepreneur or endorsing your business, or just to increase the ranking of your website. Even at this digital age people still focus on the top ten results, thus to cope up with the competition by providing organic content and using the right techniques, the information might be helpful. The emphasis was not only on how to efficiently use the SEO techniques to increase the ranking of your website just by making the user understand the working of the Search Engine System but also explaining the obstacles and making them aware of “Junk Results” which could hamper the growth of webpage. The [paper](http://infolab.stanford.edu/~backrub/google.html) by Sergey Brin and L. Page gave an insight on the working, motive and state of mind with which Google was made.

**CONCLUSION**

This paper presents an overview of search engine’s working and techniques to improve the ranking of webpages. In order to improve retrieval accuracy of Web search, we studied its architecture and tools for web-based retrieval. Our proposed approaches described in this paper contribute for indexing a target Web page more accurately and allowing each user to understand, perform more fine-grained search that satisfy their information need.

REFERENCES

[1] Broder, A. (2002) A taxonomy of web search. SIGIR Forum. 36:2. p. 3-10

[2] E. Selberg and O. Etzioni, Multi-service search and comparison using the MetaCrawler, in: Proceedings of the 4th

World Wide Web Conference (WWW4), 1995.

[3] [Search Engine Indexing](https://www.deepcrawl.com/knowledge/technical-seo-library/search-engine-indexing/)

[4] GIANNOTTI, F., NANNI, M., PEDRESCHI, D., AND SAMARITANI, F. 2003. WebCat: Automatic categorization of

Web search results. In Proceedings of the 11th Italian Symposium on Advanced Database Systems (SEBD), S. Flesca,

S. Greco, D. Sacc` a, and E. Zumpano, Eds. Rubettino Editore, 507–518.

[5] [Search Engine Indexing](https://www.deepcrawl.com/knowledge/technical-seo-library/search-engine-indexing/)

FERRAGINA, P. AND GULLI, A. 2004. The Anatomy of SnakeT: A Hierarchical Clustering Engine for Web Page

Snippets. In Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in

Databases. Lecture Notes in Computer Science, vol. 3202. Springer, 506–508

[6] [History of Search Engines](https://neolabs.ca/search-engines-history/)

[7] [The Search Result Page of Clusty - a popular clustering Search Engine. [Clustering in Search Engines]](https://www.cmi.ac.in/~tmharish/thesis/html/node45.html#:~:text=Clustering%20Search%20Engines%20unlike%20traditional,relationships%20between%20items%5B30%5D.)

[8] https://www.ecreativeworks.com/blog/what-is-google-crawling-and-indexing

[9] http://www.w3.org/People/Berners-Lee/ WorldWideWeb.html

[10] <http://en.wikipedia.org/wiki/Web_browser>

[11] <http://www.webopedia.com>.

[12] Franklin, Curt. How Internet Search Engines Work, 2002. [www.howstuffworks.com](http://www.howstuffworks.com)

[13] <http://docs.google.com>

[14] S. E. Robertson and K. S. Jones. Relevance Weighting of Search Terms. Journal of the American Society for Information Sciences, 27(3): pages 129–146, 1976.

[15] J. Rocchio. Relevance Feedback in Information Retrieval. In G. Salton, editor, The Smart Retrieval System: Experiments in Automatic Document Processing, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[16] G. Salton. The Smart Retrieval System: Experiments in Automatic Document Processing. Prentice - Hall, Englewood Cliffs, NJ, 1971.

[17] C. J. van Rijsbergen. Information Retrieval. Butterworths.

[18] <http://en.wikipedia.org/wiki/Search_engine>

[19] http://www.pewinternet.org/Reports/2012/ Search-Engine-Use2012/Summary-of- findings.aspx