VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

SEARCH ENGINE

VYHLEDÁVAČ

Ha Xuan Huy Truong

Bachelor

Supervisor: Ing. Hussam Abdulla, Ph.D.

Ostrava, 2022

VSB – Technical University of Ostrava

Faculty of Electrical Engineering and Computer Science

Department of Computer Science

Bachelor Thesis Assignment

Student:	**Ha Xuan Huy Truong**
Study Programme:	B2647 Information and Communication Technology
Study Branch:	2612R025 Computer Science and Technology
Title:	Search Engine Vyhledávač
The thesis language:	English

Description:

This search engine is developed using web annotation. It is one of the trending computer science projects where when users enter specific words or phrases in a search engine, it automatically fetches the most relevant pages that contain those keywords. The aim of this work is to describe methods for Search engine.Selected methods will be described, implemented and evaluated on suitable datasets.

The work will include:
1. Description of Search engine.
2. Description of selected algorithms.
3. Design of implementation and implementation of algorithms.
4. Experimental verification of algorithms and their comparison.

References:

[1] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to InformationRetrieval, Cambridge University Press. 2008.
[2] Ricardo Baeza-Yates , Berthier Ribeiro-Neto. Modern Information Retrieval: The Concepts andTechnology behind Search. ISBN13 9780321416919, 10 Feb 2011
[3] Bruce Croft, Donald Metzler, Trevor Strohman. Search Engines: Information Retrieval in Practice 1stEdition. ISBN-13: 978-0136072249.

Extent and terms of a thesis are specified in directions for its elaboration that are opened to the public on the web sites of the faculty.

Supervisor: **Ing. Hussam Abdulla, Ph.D.**

Date of issue: 01.09.2021

Date of submission: 30.04.2022

doc. Ing. Petr Gajdoš, Ph.D.
Head of Department

prof. Ing. Jan Platoš, Ph.D.
Dean

Abstrakt

Cílem této bakalářské práce je navrhnout a vytvořit webové stránky pomocí vyhledávačů. Stručně řečeno, můj web pomůže uživatelům najít články s recenzemi produktů podle zadaných klíčových slov. Kromě toho bude k dispozici řada podpůrných funkcí pro zvýšení uživatelského pohodlí. V první části je koncept a vliv vyhledávače; další částí je analýza struktury systému; třetí část je podrobný popis celého systému z databáze, funkcí a technologií použitých v tomto projektu; Poslední částí je realizace a uzavření projektu.

Klíčová slova

Vyhledávač, webový vyhledávač, prohledávací webový vyhledávač.

Abstract

The purpose of this bachelor thesis is to design and build a website using Search Engines. In short, my website will help users find product review articles by input keywords. In addition, there will be a number of supporting functions to enhance user convenience. In the first part is the concept and influence of Search Engine; the next part is the analysis of the system structure; the third part is a detailed description of the entire system from the database, functions and technologies used in this project; The final part is project implementation and conclusion.

Keywords

Search engine, web search engine , crawler web search engine.

Acknowledgement

First of all, I would like to sincerely thank Mr. Ing. Hussam Abdulla, Ph.D. , who has always accompanied , supported and given advice during arduous and challenging times .

I would also like to thank all the teachers who have taught me over the years.

I would also like to thank Phan Dao, MSc, Ph.D., who established a joint program company between Ton Duc Thang University and VSB TUO - Ostrava Technical University.

It is impossible not to mention Jana Bogdova, she is always dedicated to solving difficulties and providing dedicated guidance throughout the study period. I am truly grateful for that.

Finally, I would like to thank my family and friends for always accompanying me from Vietnam to the Czech Republic. I hope that all my efforts can meet everyone's expectations for me.

Table of Contents

List of Figures

List of Tables

List of source code

List of abbreviations and symbols used

SERP	- Search Engine Results Page
DMOZ	- Directory.mozilla.org
SEO	- Search Engine Optimization
LSI	- Latent Semantic Indexing
CTR	- Click-through rates
RSS	- Really Simple Syndication
XML	- eXtensible Markup Language
HTML	- Hyper Text Markup Language
CNN	- Caple News Network
URL	- Uniform Resource Locator
DNS	- Domain name system
PDF	- Portable Document
MVC	- Model-View-Control
CSS	- Cascading Style Sheets
PHP	- Hypertext Preprocessor

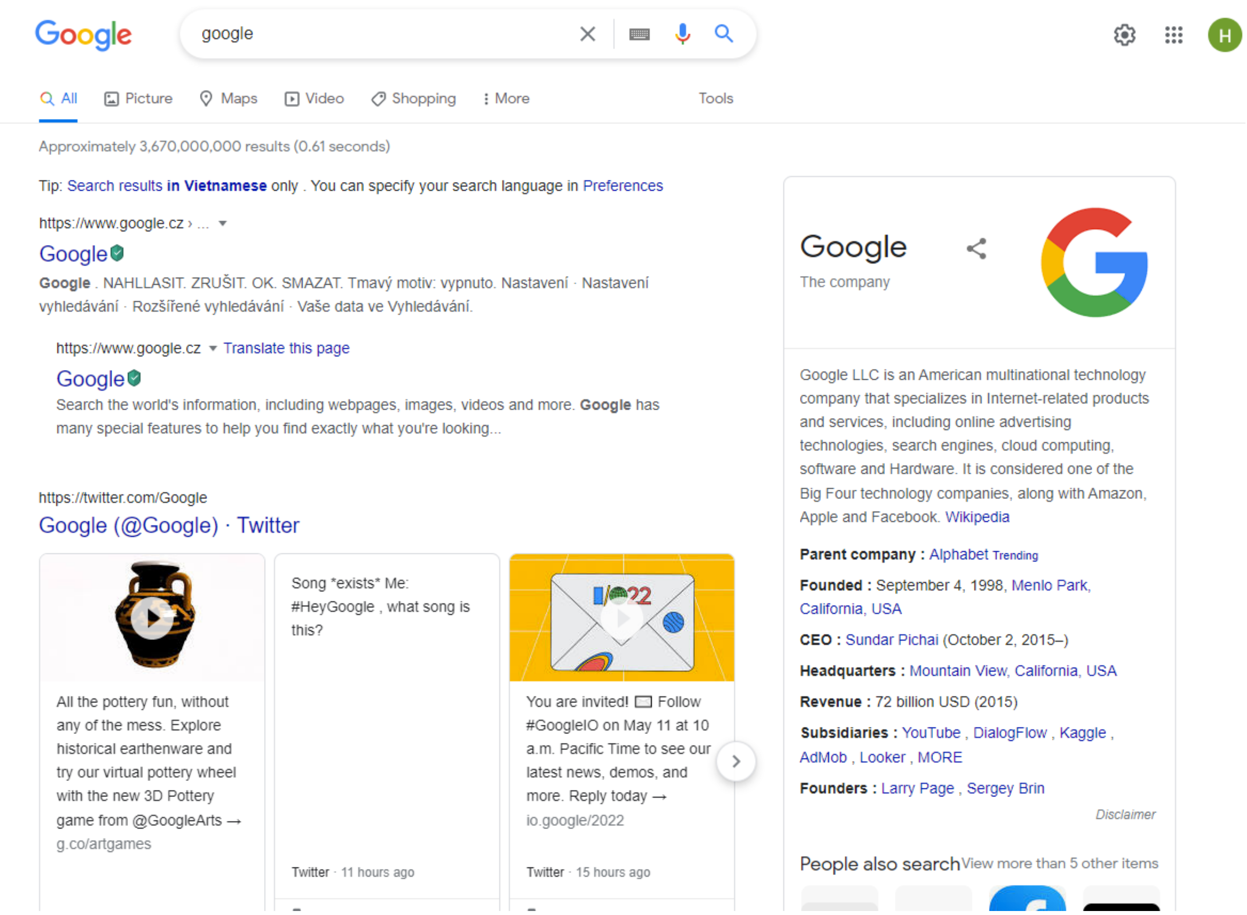
Chapter 1 Introducion

1.1 What is Search Engine ?

Simply , a search engine is a website where users type the word or phrase they want to find into the search box to see results such as web pages, images, videos, addresses, maps, documents, etc. that are relevant to what they are looking for. The words or phrases that the user enters are called keywords. The results displayed for this keyword are shown in a specific order, determined by a special algorithm of the search engine used by the user.

1.2 Search Engine Results Page

Search results are usually presented in a series of results, often called to as search engine results pages (SERPs): A Search Engine Results Page (SERP) is a web page that is displayed after a user performs a search in a search engine. The results page displays the list of results for a keyword search and allows the user to select the desired page, usually from a vertical list. Normally, there will be the following types of results displayed on the SERP: organic results ,sponsored results ,rich snippets ,featured snippets and knowledge graph.



Approximately 3,670,000,000 results (0.61 seconds)

Tip: Search results in **Vietnamese** only . You can specify your search language in Preferences

https://www.google.cz › ...

Google

Google . NAHLLASIT. ZRUŠIT. OK. SMAZAT. Tmavý motiv: vypnuto. Nastavení · Nastavení vyhledávání · Rozšířené vyhledávání · Vaše data ve Vyhledávání.

https://www.google.cz Translate this page

Google

Search the world's information, including webpages, images, videos and more. **Google** has many special features to help you find exactly what you're looking...

https://twitter.com/Google

Google (@Google) · Twitter

All the pottery fun, without any of the mess. Explore historical earthenware and try our virtual pottery wheel with the new 3D Pottery game from @GoogleArts → g.co/artgames

Song *exists* Me: #HeyGoogle , what song is this?

Twitter · 11 hours ago

You are invited! Follow #GoogleIO on May 11 at 10 a.m. Pacific Time to see our latest news, demos, and more. Reply today → io.google/2022

Twitter · 15 hours ago

Google

The company

Google LLC is an American multinational technology company that specializes in Internet-related products and services, including online advertising technologies, search engines, cloud computing, software and Hardware. It is considered one of the Big Four technology companies, along with Amazon, Apple and Facebook. Wikipedia

Parent company : Alphabet Trending

Founded : September 4, 1998, Menlo Park, California, USA

CEO : Sundar Pichai (October 2, 2015–)

Headquarters : Mountain View, California, USA

Revenue : 72 billion USD (2015)

Subsidiaries : YouTube , DialogFlow , Kaggle , AdMob , Looker , MORE

Founders : Larry Page , Sergey Brin

Disclaimer

People also search View more than 5 other items

Figure 1: Search Engine Result Page of Google

1.2.1 SEO

1.2.1.1 What is SEO ?

SEO is an acronym for Search Engine Optimization, is a process of improving the ranking of a website on search engines so that users can find a website more easily on the SERP.In search engine optimization, you must keep in mind that each search engine has a different algorithm. Depending on the search engine you are targeting, you need to carefully study the way of keyword ranking and website ranking of the search engine to develop an effective SEO strategy.

1.2.1.2 What is the importance of SEO for website?

Nowadays, the Internet has become a place where all the necessary information of users is stored. The strong development of the Internet leads to the fact that the behavior of users is also changing. Instead of direct purchasing as in the previous traditional way, users have made more purchases on the Internet.Therefore, the importance of SEO is undisputed.

The benefits that SEO brings:

- Increasing the number of potential customers.

- Increasing brand awareness.
- Optimizing customer access costs.
- Helping you understand your customers better.

Notes when doing SEO for website:

- Optimize the crawlability of Search Engine websites: To optimize the data collection process, you need to pay attention to the following information such as server,hosting; HTTP status code; Robot meta tags ; HTML code of web content.
- Keyword research: is to find the set of keywords that need to be optimized and support the process of building a content strategy. When researching keywords, you need to pay attention to information related to: Search intent, LSI keyword and Long tail keyword.
- Website structure optimization: is how you organize website content. Site structure is concerned with how this content is grouped, linked, and presented to the reader. If you have a good site structure, it will be easier for users to find and also better index your URLs.
- Building standard SEO content:Articles with quality content will be appreciated.
- Onpage optimization: can simply be understood as all the things you optimize on the page. Onpage SEO factors include technical factors (e.g. website code quality and website speed) and other factors. factors content-related aspects, like the quality of the content on your site.
- Off-page optimization:are optimizations outside the website. These jobs can include link building or social optimization and brand reputation building.
- Analyze and measure SEO effectiveness: you need to use tools to be able to analyze and measure the effectiveness of the SEO process. Thanks to these analysis tools we can know important data information such as the number of visitors; demographics of visitors ; time on-site; the number of pages per session (Page View) and many other metrics.
- Optimizing for mobile devices: Optimizing your website to be more mobile-friendly has become one of the most important criteria.
- Optimizing User Experience: To optimize user experience, optimization should be kept in mind: Dwell time (time spent by the user on the page during each session); search intent ; CTR (click-through rate) and bounce rate.

1.2.1.3 Some popular types of SEO

- Map SEO (Local Search Optimization)
- Image SEO (Image Optimization)
- Video SEO (Video Optimization)
- Voice search (Optimize voice search)

1.3 Why we need Search Engine ?

In today's life, everything is more and more developed and modern. Most people use the internet as a tool for life. They will likely be searching for something with hot keywords to shopping; find documents,

information or entertainment, etc. The more detailed the hotkeys, the more accurate information we can find. This will save time and be convenient for everyone.
In summary, using Search Engines brings many advantages:

- Time Savings. A search engine saves you time in two ways: by eliminating the need to find information manually, and by performing searches at high speeds.
- Relevance. When a search engine scans a website, it scores the content for relevance to particular search words.
- Free Access.
- Comprehensive.
- Advanced Search.

To learn more about the importance of search engines in some common areas of everyday life:

1.3.1 Using Search Engines To Shop

In a time when the internet was not popular, we would find logos of many companies when we were on the street, in newspapers or on TV.
Let's imagine how one person can remember all the information, product name, price or brand of the product. Then we will search by the hot keys we want, for example: find the car we like through the name of the car manufacturer, the name of the phone ...
During the Covid-19 pandemic, it will be very difficult for us to shop in person. This makes online shopping even more essential. Therefore, search engine becomes more and more important.

1.3.2 Using Search Engines To Research or Study

As a student , information retrieval , learning material or data . Actually everything would be very difficult if we had to read every book or newspaper to find the necessary information . That will take a lot of time. Instead, we will enter the information we want to search through popular search sites such as Google, Yahoo, Bing, Baidu, ...
But that the informations is provided by many other people around the world so we need to study it carefully. For examples, we will read in many reputable sites or popular books.

1.3.3 Using Search Engines To Find Entertainment

Besides research and shopping, entertainment is also affected. After going through a stressful time of work and study, many people will use search engines as a tool for entertainmentThey look up things like videos, movies, games, and social networking sites,... then easily download to their computer or smartphone.
If you want to find videos, Youtube is probably a popular choice. People can easily find a lot of videos with various entertainment content about music, movies, sports, advertising or fashion shows, ... others like to watch food review videos. , travel or funny videos, ...Totally,Youtube brings videos with high image quality, multi-content and fast that is a good idea for relax.

Chapter 2 Analyzing the architecture of a Search Engine

In this section, we will analyze the overall structure of the search engine. An architecture is designed to ensure that a system meets the requirements or goals of the application. The two main goals of a search engine are:

- Effectiveness (quality): we want to be able to retrieve the most relevant documents for a query.
- Efficiency (speed): We want to process users' queries as quickly as possible.

2.1 Basic building blocks

2.1.1 Search Engine Components

Search engine will consist of three basic main components, listed below:

- Web crawler: It is also known as spider or bots. It is considered a piece of software used to collect data from websites.

- Database: This will be the place to store all the website's information data.

- Search Interfaces: This is an interface component to help users connect to the database, from which users can search data through the database.

2.1.2 How is Search Engine Working ?

The first thing, we need a software called spider to be able to collect information from millions of websites that exist on the network, Starting from any website, spider will crept into every corner of that page and visit each link on the page in turn. Then bookmark the previously visited links and connect the linked pages to the original page like creating a chain linking two pages together. Simply from an initial web, Spider can connect many other sites to form a dense web like a real spider's web. The spider will build its list, a process known as a web crawler. Besides, the indexing process happens immediately and parallel to the crawling step above. Index will do the saving of all data collected on a web page. The data after saving will be encrypted as a text file so that it can be saved with the lowest size to help extract the fastest search results. After the data is successfully encrypted, it will be further analyzed and indexed and saved in the database, helping Search Engine not have to search for information on each website every time someone searches for a keyword. Upon receiving a user's search request, the Search Engine will take the information stored in the database, sort the results found, and display a list of answers to us through the search interface. The Search Engines rely on two criteria to evaluate the order of search results: relevance and popularity. Search results related to your request are prioritized, then the popularity of each result is considered. However, different Search Engines have a different way of assessing relevance and popularity. This is the difference in the search algorithm of each search engine. This retrieval process is also the main object of the intervention of SEO tricks. Through SEO, we make the Search Engine algorithm appreciate a certain search result and change its ranking in the final list of search results.

2.2 Architecture

In more detail, the search engine architecture consists of two basic layers listed below:

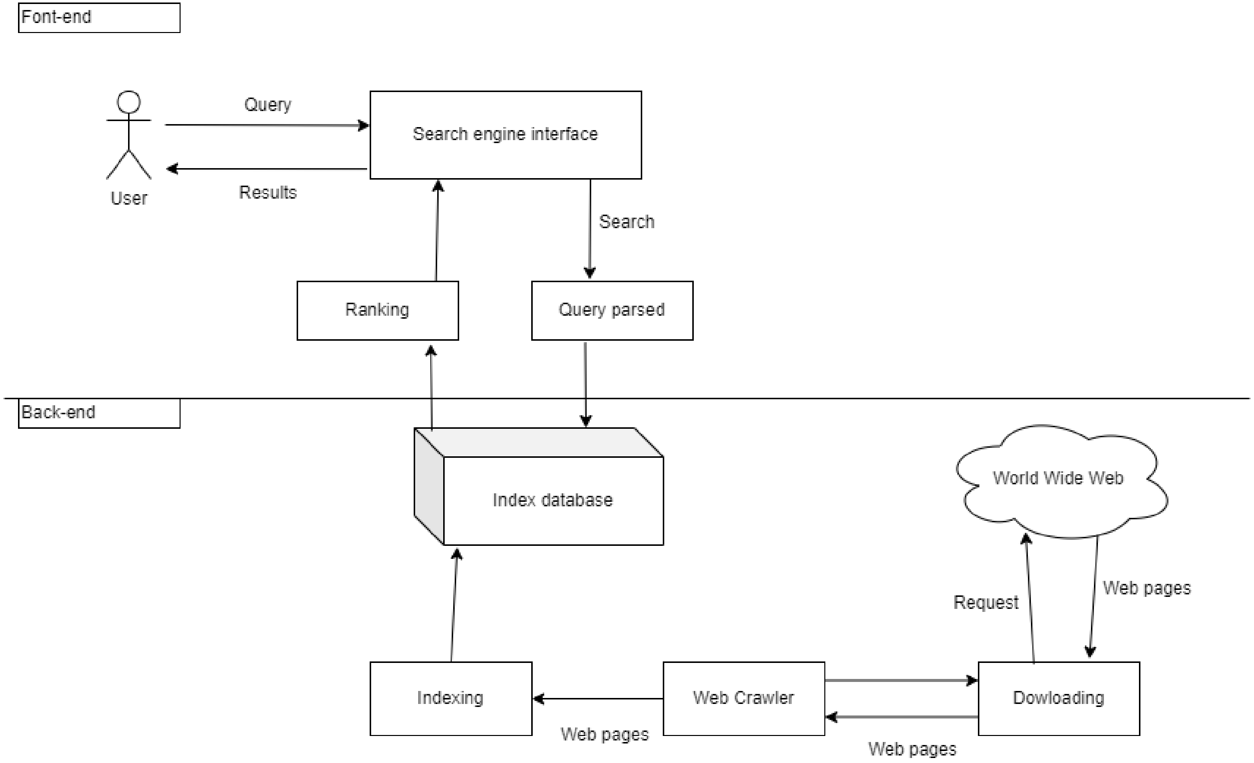


Figure 2:Architecture of Search Engine

2.2.1 Search Engine Processing

The indexing process and the query process are two major functions supported by search engine components.

2.2.1.1 The indexing process

The indexing process includes the following three tasks:

- Text acquisition: it identifies and stores documents for indexing.

- Text transformation: The document is converted into index terms or features.

- Index creation: The index terms generated by text conversion are used to create data structures that enable fast searching.

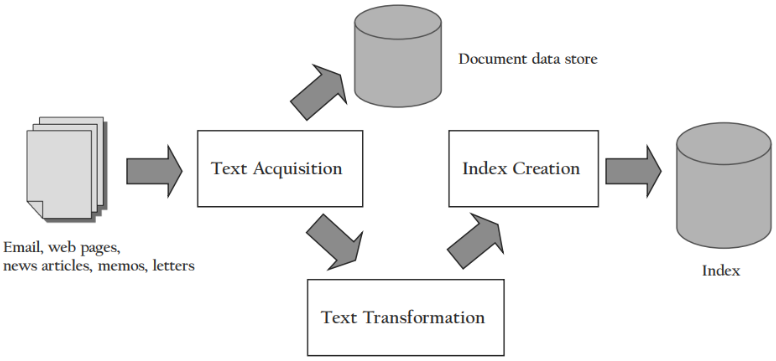


Figure 3 :The index process[4]

The indexing process creates the structures that allow for searching, and the query process uses those structures as well as a user's query to generate a ranked list of documents. The "building blocks" of the indexing process are illustrated in Figure 3[4].

- The text acquisition component's job is to find and make the documents that will be searched available. Although it may be possible to use an existing collection in some cases, text acquisition is more likely to require crawling or scanning the Web, a corporate intranet, a desktop, or other sources of information to build a collection. The text acquisition component creates a document data store, which contains the text and metadata for all of the documents, in addition to passing them on to the next component in the indexing process. The document type (e.g., email or web page), document structure, and other features, such as document length, are all examples of metadata.

- Documents are transformed into index terms or features using the text transformation component. The parts of a document that are stored in the index and used in searches are referred to as index terms. A word is the most basic index term, but not all words are searchable. In the field of machine learning, a "feature" refers to a portion of a text document used to represent its content, which also describes an index term. Phrases, people's names, dates, and web page links are all examples of index terms or features. The term "terms" is often used to describe index terms. "Terms" is used to refer to index terms. The index vocabulary is the set of all the terms that are indexed for a document collection.

- The index creation component uses the text transformation component's output to create indexes or data structures that allow for quick searching. Because many search applications have a large number of documents, index creation must be efficient in terms of both time and space. When new documents are acquired, indexes must be able to be updated quickly. By far the most common type of index used by search engines is inverted indexes, also known as inverted files. Simply put, an inverted index is a list of all the documents that contain the index term for each index term. It is inverted in the sense that it is the polar opposite of a document file that lists the index terms that each document contains. Inverted indexes come in a variety of shapes and sizes, and the type of index used is one of the most important aspects of a search engine.

2.2.1.2 The query process

The query process consists of the three tasks listed below:

- User interaction: it allows for the creation and refinement of the user query, as well as the display of the results.
- Ranking: It creates a ranked list of documents using queries and indexes.
- Evaluation: it keeps track of and measures effectiveness and efficiency. It is carried out offline.

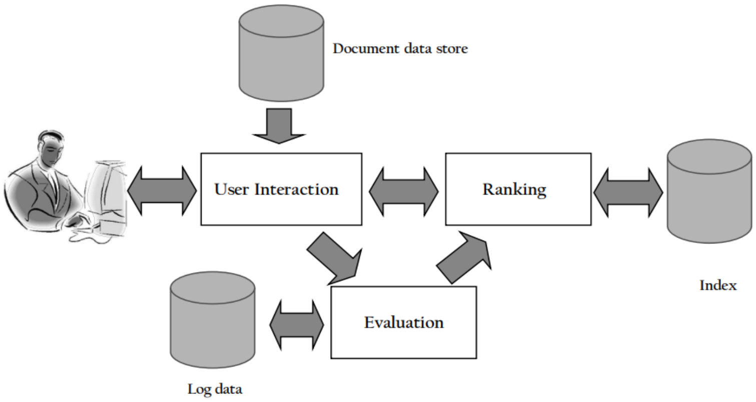


Figure 4: The query process[4]

The building blocks of the query process are depicted in Figure 4[4].

- The user interaction component acts as a conduit between the person searching and the search engine. Accepting the user's query and transforming it into index terms is one of this component's responsibilities. Another task is to organize the search engine's ranked list of documents into the results displayed to the user. This can include things like creating the snippets that are used to summarize documents. One of the sources of data used to generate the results is the document data store. Finally, this component includes a number of techniques for fine-tuning the query so that it more accurately reflects the information requirement.
- The search engine's core component is the ranking component. It takes the transformed query from the user interaction component and uses retrieval model scores to create a ranked list of documents. Because many queries may need to be processed in a short period of time, ranking

must be both efficient and effective, as the quality of the ranking determines whether the search engine achieves its goal of finding relevant information. The effectiveness of ranking is determined by the retrieval model, and the efficiency of ranking is determined by the indexes.

- The evaluation component's job is to assess and track the effectiveness and efficiency of the program. Using log data to record and analyze user behavior is an important part of that. The ranking component is fine-tuned and improved based on the evaluation results. Apart from logging user and system data, the majority of the evaluation component is not included in the online search engine. Although evaluation is primarily an offline task, it is an essential component of any search application.

2.2.2 Web crawler

Web crawlers are websites that collect, filter, and search data on the internet from World Wide Web sites that are accessible over the network.A web crawler (also known as a web spider/web robot, ants, automatic indexers, bots, and worms) will extract information that matches the user's query with the original link for easy access by the user... Web crawling, also known as spidering, is the process.

Everything on the website will eventually be found and spidered from another website, which is simple to understand. On multiple servers, search engines can simultaneously run thousands of web crawlers. When a web crawler visits your site, it will download all of the content and upload it to its database right away. Your website's content is stored in the search engine index, which is a massive database of keywords that have appeared on a variety of websites. As a result, getting your website to the top of the search results will be a fierce battle.
The first step is to populate your website with content, preferably of high quality.Only a robots.txt file can tell the spider which parts of your website it should crawl and which parts it should ignore. We will learn more about it in chapter 3.

2.2.3 Storing the documents

The majority of search engines require a storage location for documents. To create document snippets for each search result, quick access to the document text is required. Without having to click on a link, these snippets of text give the user an idea of what's inside the retrieved document.
Even if snippets aren't required, it's still a good idea to keep a copy of each document. In terms of both CPU and network load, document crawling can be costly. It's more practical to keep copies of the documents on hand rather than trying to retrieve them again the next time you need to create an index. Keeping old documents allows you to save bandwidth by using HEAD requests in your crawler or crawling only a subset of the pages in your index.
Finally, document storage systems can be used to begin the process of extracting information. Web search engines extract anchor text from links to store with target web documents, which is the most common type of information extraction. Other types of extraction, such as identifying names of people or places in documents, are possible. It's worth noting that if the search application uses information extraction, the document storage system should be able to modify the document data. We'll now go over some of the fundamental requirements for a document storage system, such as random access, compression, and updating, as well as the pros and cons of using a database system versus a customized storage system like Google's BigTable.

2.2.3.1 Database system

A database is an excellent place to store documents for many applications. A database handles the difficult details of storing small pieces of data, such as web pages, and makes updating them easy later. Most databases also function as a network server, allowing users to access documents over a network. This could allow a single computer to serve documents for snippets while several other computers handle queries. Databases often include useful import and analysis tools that make managing the document collection easier.

Many companies that run web search engines store documents in traditional relational databases. The sheer volume of document data is one issue that can overwhelm traditional database systems. Database vendors also assume that database servers will use the most expensive disk systems available, which is unrealistic given the size of the collection.

2.2.3.2 Random Access

The document store must support random access in order to retrieve documents quickly in order to compute a snippet for a search result. In comparison to a full relational database, only a simple lookup criterion is required. We want a data store that allows us to request a document's content based on its URL. Hashing is the simplest way to handle this type of lookup. By applying a hash function to the URL, we can generate a number that we can use to locate the data. The hash function can tell us which file contains the document in small installations. The hash function tells us which server contains the document in larger installations. A B-Tree or sorted data structure can be used to find the offset of the document data within the file once the document location has been narrowed down to a single file.

2.2.3.3 Compression

Compression works best with large blocks of data, so it's ideal for large files containing many documents. However, compressing the entire file as a single block is not always a good idea. Because most compression methods don't allow for random access, each block must be decompressed one at a time. If you want random access to the data, compressing in smaller blocks, perhaps one block per document or one block for a few documents, is a better option. Small blocks improve request latency while reducing compression ratios (the amount of space saved).

2.2.3.4 Updating

It makes sense to update the document store as new versions of documents arrive from the crawler. The alternative is to create an entirely new document store by merging the crawler's new, changed documents with document data from the old document store for documents that did not change. This merging process will be much more expensive than updating the data in place if the document data does not change much. An example link with anchor text:

```
<a href="http://example.com" >Example website</a>
```

Handling anchor text is another important reason to support update. An example of anchor text in an HTML link tag can be seen in the example. The HTML code in the example will appear in the web browser as a link with the text Example website, which will direct the user to http://example.com when clicked. Anchor text is useful because it gives a quick overview of what the target page is about. We may believe that the summary is unbiased if the link comes from a different website, which also helps us rank documents. Anchor text collection is difficult because the anchor text must be linked to the target page. Using a data store that supports updates is a simple way to go about it. When we find a document with anchor text, we look for the target page's record and update the anchor text portion of the record. When it's time to index the document, all of the anchor text is in one place and ready to go.

2.2.3.5 BigTable

For most of Google's backend and online applications/products, Google BigTable is a multidimensional, distributed, and multidimensional data storage mechanism based on the company's proprietary storage technologies. For extremely large database infrastructures, it provides a scalable data architecture. Google BigTable is a sorted and continuous map. Each string in the map has a row, a column (of some kind), and a timestamp value that is used to index the data. A web page's data series looks like this:

- The row name is used to store the reverse URL (com.google.www).
- The content column stores the content of a Web page.
- Any anchor text or page reference content is saved as anchor content.
- Timestamps are used to organize multiple versions of a page and provide the exact time when data was stored.

BigTable uses immutable (unchangeable) files to store its data. The data in a file is never changed once it is written to a BigTable file. This also aids in the recovery of failures. Because only some of the outstanding writes completed before the computer crashed, failure recovery in relational database systems necessitates a complex series of operations to ensure that files are not corrupted. A file in BigTable is either incomplete (in which case it can be thrown away and recreated from other BigTable files and the transaction log) or complete (in which case it cannot be thrown away and must be recreated from other BigTable files and the transaction log). The most recent data is stored in RAM, while older data is stored in a series of files to allow for table updates. The files are periodically merged to reduce the total number of disk files.

2.2.4 Search interface

Employees search a data source using a search interface, which is a graphical user interface. A search interface can be created for any device, from a phone to a computer. Search interface also known as UI for user search.The concept of interface is fairly straightforward when it comes to computers and user interfaces. The following will provide information on these two topics:

- In computers:Communication ports present in the network state are commonly referred to as interfaces. Information can be shared between computers using these communication ports.

- In UI (User Interface):In programming, an interface is a collection of objects that can be selected or reset by the user. Both Android and Windows operating systems support the user interface. Only when programming changes will the Interface take on a variety of forms.

2.3 Different Types of Search Engines

Search engines are classified into the following four categories based on how it works:

Crawler based search engines
Human powered directories
Hybrid search engines
Other special search engines

We will analyze each method in more detail in the next section.

Chapter 3 Analyzing Search Engine methods

3.1 Crawler Based Search Engines

All crawler based search engines use a crawler or bot or spider for crawling and indexing new content to the search database. There are four basic steps, every crawler based search engines follow before displaying any sites in the search results.

- Crawling
- Indexing
- Calculating Relevancy
- Retrieving the Result

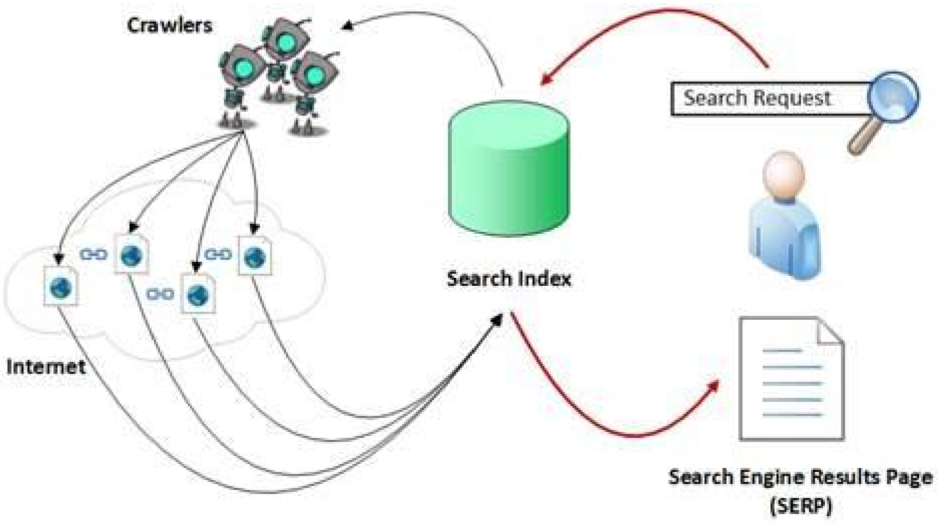


Figure 5:How Crawler Based Search Engines work.[4]

In my opinion, Crawl Based Search Engine would be the most appropriate method because of the popularity and ease of use that brings to the users. We will learn in detail how each of the above steps works:

3.1.1 Crawling

Search engines crawl the entire web in order to retrieve the available web pages. The crawling of the entire web is performed by a piece of software known as a crawler, bot, or spider. It is capable of recognizing and collecting HTML pages as well as other document types such as PDF, PowerPoint, Word, and Excel. It uses conversion software to extract text information from these documents, resulting in indexable and searchable text information. Crawling frequency is determined by the search engine, and it may take several days between crawls. This is why your old or deleted page content may appear in search results from time to time. Once the search engines crawl your site again, the new updated content will appear in the search results. The architecture of a large-scale crawler we previously developed to explore and download billions of web pages is depicted in Figure 6.

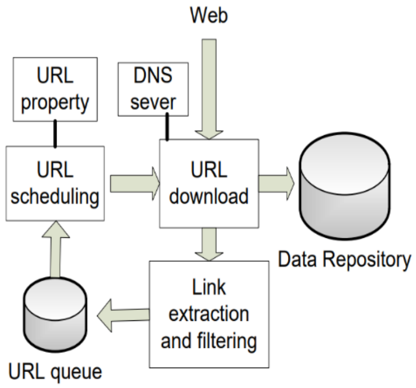


Figure 6:Architecture of a web crawler.[9]

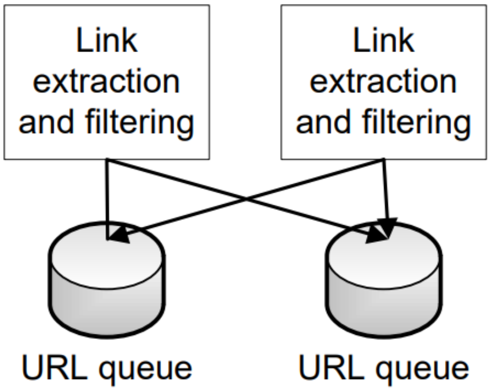


Figure 7:Outgoing link extraction and URL distribution in collaboration.[9]

- **The URL downloader retrieves web pages in accordance with the HTTP protocol**: A crawler typically uses a thread to handle an HTTP connection and runs multiple concurrent threads to consume all available download bandwidth. Because downloading a page from a site can have a high latency, a crawler machine may have a large number of HTTP requests outstanding. This increases system resource requirements because each machine must keep track of all open connections.

- **A DNS server resolves URL IP addresses**: Because there is a high volume of DNS lookups for URL hosts, additional caching support can be beneficial.

- **The URL extraction module parses downloaded pages and finds outgoing links**: Some URLs identified as low quality or duplicate may be filtered. It saves the extracted outgoing URLs to the URL queue after filtering. Because crawlers typically run their services on a group of machines, URLs extracted from one machine can be served to URL queues in multiple machines, as shown in Figure 7.

- **URL Scheduling controls which URLs are visited in the next crawl:** Small sites can be crawled quickly due to their small size, whereas large sites take a long time to access. The URL Scheduler regularly analyzes the properties of queued URLs, prioritizes them, and selects a subset of URLs for URL download to achieve the following goals: 1) The major sites are fully crawled; 2) URLs are re-crawled on a regular basis to ensure freshness; 3) re-visits of recently crawled URLs are prohibited to avoid repeated visits to the same URLs; 4) Duplicate or low-quality pages are not crawled or are crawled with low priority.

- **The URL property service manages URL properties**: There is resource contention to accomplish these goals concurrently, and the URL scheduler makes a sound decision based on URL properties such as the last successful crawling time and unsuccessful data fetch attempts. Because each crawling machine has a large number of data accesses in parallel at each second, and there are a large number of URLs to be crawled, accessing URL property information can become a crawling bottleneck. As a result, the design of an efficient data structure and crawler architecture is critical.

A number of factors contribute to the complexity of crawler management:

- **Politeness policies**: A website frequently contains a large number of pages. A crawler can consume a significant amount of computing and bandwidth resource from this site in order to fetch these pages quickly, which can disrupt the normal operation of such a site. A crawler must follow a politeness policy to limit its crawl rate in order to reduce the negative impact on the site's resources. A crawler, for example, should not make more than one request to the same host at the same time, and it should possibly add a time delay between two consecutive requests to the same host.

- **Robots exclusion protocol**: A web administrator can exclude their pages from a search engine's index by placing the /robots.txt file in the web site's top directory. This file specifies the crawler's visit restriction for a search engine [12]. A line in example.com/robots.txt that says "Disallow: /tmp/" means that visiting example.com/tmp is not permitted. A crawler must check the robots.txt file from the corresponding web host when retrieving a page from a site. A crawler may save a copy of the robot.txt file in its local cache system with a periodic update to avoid repeated access to the robot.txt file from the same host.

- **Crawling abnormality**: Some sites respond slowly or even fail to respond to a crawler's request. Some websites may provide crawlers with incorrect or broken results. A crawler must be resilient to a website's failure or slowness and be able to retry at a later time. Crawling traps are another type of crawling abnormality. A website keeps a crawler busy in this case by dynamically feeding an infinite number of useless web pages. For example, an infinite number of calendar pages can be generated dynamically and made available via a site's "Next day" button. A site-oriented similar content analysis can detect such a trap.

- **Sitemap and hidden web**: Major search engine crawlers can effectively collect surface-level web pages, but there is a vast amount of hidden deep web information that search engines cannot access. A crawler cannot find any page that is not explicitly pointed to by other web pages using hyperlink analysis. Pages that are dynamically generated and require HTML-form or Javascript submission are frequently invisible. Automatic form filling can explore the deep web [10], but it is still difficult to achieve high accuracy or coverage. An alternative method is to use the Sitemaps protocol [11], which allows webmasters to notify search engines about a list of pages on their sites that are available for crawling.A Sitemap is an XML file that contains a list of URLs for a website as well as additional metadata about each URL. As a result, webmasters can promote deep URLs that are not even hyperlinked from the surface web.

The freshness and content coverage of a crawled data collection are important for a search engine's quality. The challenge for a crawler is to quickly discover newly created or updated web pages. A crawler can return to known URLs frequently to see if their content has changed. The historical change frequency of a page and its site can help a crawler determine how frequently it should revisit this page. While crawling sites can discover new pages on a regular basis, leveraging information published by a site, such as RSS feeds, can accelerate this process. An   feed is an XML-based file that contains a structured list of newly added pages to a website. The RSS feeds are frequently associated with blogs and other publishers of information. CNN, for example, offers RSS feeds for news in various categories. Amazon publishes RSS feeds with tags for popular products. There are additional web signals that can be mined to aid in the discovery of new URLs, particularly for hot or new topics. This includes detecting new URLs in news, blogs, search logs, and social networking sites.

3.1.2 Indexing

After crawling, the next stage is indexing, which is the act of determining the words and phrases that best describe the website. The words that have been identified are referred to as keywords, and the page is allocated to the keywords that have been identified. When the crawler does not grasp the meaning of your page, it may cause your site to rank lower in the search results. You must optimize your pages for search engine crawlers in order to ensure that the material is easily readable. Once the crawlers have identified the relevant keywords, your page will be allocated to those keywords and will appear high in search results.

3.1.3 Calculating Relevancy

The search engine compares the search string in the search request to the indexed pages in the database. Because it is likely that the search string appears on more than one page, the search engine begins calculating the relevance of each page in its index to the search string.
Relevancy can be calculated using a variety of algorithms. Each of these algorithms assigns different relative weights to common factors such as keyword density, links, and meta tags. That is why different search engines return different search results pages for the same search string. It is common knowledge that all major search engines change their algorithms on a regular basis. If you want to keep your site at the top, you must also adapt your pages to the latest changes.

3.1.3.1 Ranking

Because most users only look at the top results, ranking is the most visible aspect of a search engine. Ranking can be done centralized or hierarchically, with each subsystem selecting the best results first before submitting them to the next round of ranking. Document relevance scores are calculated by utilizing their text and authority features, which are commonly referred to as ranking signals. We begin by providing an overview of query-dependent and query-independent ranking signals, followed by a description of feature analysis using text, link, and click data. Then we will discuss about score aggregation using these signals.

3.1.3.1.1 Ranking signals

There are four aspects of ranking that must be considered when matching a user's intent. 1) Relevancy. The top documents must be relevant to the user's query. 2) Authoritativeness. Users typically prefer high-quality content because they rely on trustworthy information to learn or make a decision. 3) Freshness. Some queries are time sensitive, and the most recent information is preferable for such queries. 4) Preference. Personal or geographical preferences can also influence answer selection. Rank signals are intended to reveal those aspects from the query or document's perspective, and we summarize two types of ranking signals below.

Query-dependent ranking signals:

- **Document text** :This feature compares the similarity of query words to what appears in a document. A web document can be divided into multiple text sections, such as the title and body, and query words that match in different sections have different weights. If query terms are highlighted in a document, additional weighting can be considered. The proximity, positions, and frequency with which query words are matched in a document are important indicators of relevance. Some text features in a web document are hidden from users, such as the description and keywords tags in an HTML document. Spammers frequently use such a feature to present irrelevant keywords; therefore, such features must be used with caution.

- **Anchor text**: The anchor text of a web page appears in the hyperlinks of other web documents that refer to it. Many hyperlinks may point to the same web page, and some of the anchor text may be meaningless or of poor quality. The similarity of anchor text and the popularity of source web pages can be used to determine the quality of anchor text.

- **URL text**: The text of a URL, particularly when it appears within or close to the host name, may occasionally reveal the semantic of its content. For example, the keyword "download" in http://www.microsoft.com/enus/download/ is an important keyword for matching query "Microsoft download."

- **Historical click-through data**: When users search for and browse matched results, a search engine collects user behavior data. These URLs that a user clicks after performing a search query are frequently relevant to that query.

- **Query classification and preference**: To select preferred documents, implicit query features can be extracted during classification. The language of a query, for example, prioritizes the type of web documents to be matched. The geographic origin of a query may cause the ranking system to select more content from this region. If a query matches a trending news topic, fresh content that is up to date is preferred.

- **Link citation** :This feature makes use of web hyperlink connectivity to find pages with text that matches a given query. A page that is cited by a large number of other matched web pages can be a good indicator of authority for this query.

Query-independent signals:

- **Link popularity**: This feature examines how pages link to one another and then uses that data to determine the importance of each page on the web. When a page is cited by a large number of other pages, it is often ranked higher.

- **Documentation classification and spam analysis**: A classification of web documents can provide information about matching user preferences. Users, for example, are more likely to seek content in their native language than in other languages. Users may also prefer file types such as HTML and Microsoft Word. A spam score is also a useful quality indicator because spam analysis results in a rating of page quality.

- **URL Preference**: users prefer a short URL and prefer a static URL over a dynamically generated URL. Dynamic generated URLs frequently change their content, and their average quality is frequently lower than that of static URLs.

- **Site or page properties**: Some additional information on a web page or its site can be an indication of the quality of the web page. For example, the frequency of user visits, the length of the visit, and the web host's serving speed or dependability.

3.1.3.1.2 Text analysis

The goal of text analysis is to determine how relevant a document is based on how query words appear in it. For advanced retrieval semantics, a search engine can use Boolean and phrase operators. "Retrieval" or "Search," for example. We focus on text analysis of a search query with conjunctive semantics because most people use simple and short queries.
When creating a formula to measure the text closeness of document-query matching, there are three issues to consider:

- What are the fundamental units of data used to match a query to a document? After removing HTML tags, a document is typically tokenized as a series of words, with each word serving as a basic term for matching. As a result, a document can be thought of as a collection of words. On the other hand, the order and proximity of terms are critical. "Paris Hilton," for example, is the name of an actress, while "Hilton Paris" could refer to the Hilton hotel in Paris.

- What is the distance between search words that appear in a document? When search words appear close together in a document, it is often assumed that the document has a high relevancy. An n-gram term, which is defined as a consecutive sequence of n words in a document, can be used to capture word proximity. A "unigram" is a size 1 n-gram, a "bigram" is a size 2, and a "trigram" is a size 3 n-gram.
  For a simple example we have the string "Hello world!", parsed into bigram:
  "Hello world!" => {"He", "ll", "o ", "wo", "rl", "d!"}

- What is the weighting of terms? One approach is to use a simple weighting system based on the order in which they appear. A query term will be given credit if it appears in the body of a document. It receives weighted credit if it appears in the title. TF-IDF weighting (Term Frequency and Inverse Document Frequency) is another method that uses frequency information. When a term appears multiple times in a document, its importance rises in this scheme. When a term appears multiple times in multiple documents, its weight decreases because it becomes less effective at distinguishing between relevant and irrelevant documents. TFIDF can be calculated using the following formula:

  $$\mathrm{tfidf}(t, d) = \mathrm{tf}(t, d) * \log n/\mathrm{df}_t$$

  where n is the total number of documents in the collection, tf(t, d) is the number of times term t appears in document d, and df_t is the total number of times term t appears in the collection. Because the denominator df_t could be zero, it's usually adjusted to $1 + \mathrm{df}_t$. In some cases, tf(t, d) is normalized against the length of the document d. In the literature, variations of the TF-IDF weighting have been proposed, such as Okapi BM25 [13].

- How can aggregating weights from multiple query terms be used to represent the closeness of matching? The cosine-based similarity function is one model. Allow document d to contain a set of terms, each with a weight of w_i, and query q to contain a set of query terms, each with a weight of qw_i. These weights are normalized, and the total text score is calculated:

  $$\text{Score}(d, q) = \Sigma_{ti \in q}\, w_i * qw_i \,.$$

  There are other methods for combining weights. Because text features can be represented in a variety of ways to capture the appearance and proximity of search terms, a weighted summation function or a machine learning model can be used to aggregate them.

3.1.3.1.3 Link analysis

In this section, I will introduce the following two algorithms:

- **PageRank Algorithm:** The earlier PageRank algorithm [58] measured the importance of a web document based on its presence in a large web graph as a query-independent ranking signal. Each node in this graph represents a web page, and each edge represents a link between two pages using the HTML hyperlink notation. The idea behind this link analysis is that pages with a lot of links are more important than pages with few links, but not all links are equal in importance. A page with a link from a well-known site may be ranked higher than pages with more links from less well-known sources.
  Let n represent the number of pages in a web graph and x represent a single page. The PageRank model can be expressed in a simplified form as follows:

  $$R(x) = \frac{\lambda}{n} + (1-\lambda)\Sigma_{y \in P(x)} \frac{R(y)}{|S(y)|}$$

  where S(x) is the set of pages which x points to, and P(x) is the set of pages that point to x. Iterative PageRank computation is possible. Every page has a value of 1/n when it is first created. Then, for each page, a -linear combination of a rank value 1 /n and additional rank credit propagated from its predecessors in the graph will be assigned. The vector R() is then normalized. This process is repeated until the rank values of all nodes have converged.

  The in-degree of a web page can be used to approximate its link popularity score, but this score is more susceptible to link spam. That's because it's relatively simple to create thousands of Web pages pointing to a single page. When Google first launched, the PageRank algorithm was seen as the key to distinguishing Google from other search engines. Various modifications to improve relevancy or speed up computation have been proposed in the literature. See, for example, [14, 15].

- **HITS algorithm:** To determine the authoritativeness of a web page relevant to a query, the HITS algorithm [16] uses query-dependent scoring. HITS differs from PageRank in that it calculates link scores within a subgraph of a web graph that is derived from a user query. Every page receives two scores as a result of the HITS algorithm. The authority score, which represents the authoritativeness of a page relevant to the subject of a user query, is the first score. The hub score, which represents its rank value in serving as a compilation of relevant information, is the second score. Let A(x) be the authority score of a page, and H be the hub score of this page (x).

$$R(x) = \frac{\lambda}{n} + (1 - \lambda)\Sigma_{y \in P(x)} \frac{R(y)}{|S(y)|}$$

where P(x) contains the predecessors of page x in this graph, and S(x) contains the successors of x.

These scores are calculated in an iterative manner. Following an initial score assignment, all pages update their hub and authorities using the formula described above. At each iteration, these values are normalized, and the process is repeated until all pages have a converged value. The above procedure computes the principal Eigen vector of the matrix C^tC and CC^t, where C is the connectivity matrix based on the query graph's link structure, and C^t is the transpose of C. The most difficult aspect of using HITS is the cost of online computation. There are several extensions or follow-up projects to improve PageRank and HITS' seminal work. Topic distillation [17] and multi-community hub-authority [18, 19] are examples of this. Teoma/Ask [20] developed a fast algorithm for computing hub authority scores using sparse matrix approximation.

3.1.3.1.4 Click analysis

Users interact with a search engine by clicking on a query's search results, browsing them, reformulating their queries, and doing more browsing. A search engine records such interactions, and the click data is represented as a tuple (t, u, q, c) indicating that user u clicked result c for query q at time t. While click data can be noisy, URLs that are clicked by users are likely to provide some information that can be used to determine relevance. If a document has attracted the most previous users for the same query, it will be ranked highly. For each user, a log dataset can be divided into a series of search sessions. Each user session consists of a series of queries and results chosen by the user. Each user session contains a set of queries and results that this user has chosen. Each user session contains a set of queries and results that the user has chosen. Session analysis revealed the following correlations between queries and URL results.

- Query-to-pick query (Q2P): A query is linked to a pick in a Q2P correlation. A correlation candidate is created when a search engine returns a result in response to a query and the user selects it. This Q2P correlation becomes more plausible when multiple independent users make the same connection.

  After a user enters a query, picks recorded for a different query within the same user session can be associated with the initial query as an extended association. Another benefit is that different pages visited as a result of a URL selection may be linked to the original query. The two extensions above represent the fact that a user frequently reformulates a query or browses beyond a search link to find relevant content. Such an approach is prone to data errors, necessitating the use of additional noise filtering techniques. For example, query similarity from one pick to the next in a session, the order and depth of query rewriting or browsing steps, dwelling time, and the first and last choice of browsing can all be used to apply a restriction.
- Queries-to-queries (Q2Q). This link connects queries made during a user session. The confidence in such an association is determined by a number of factors, including query similarity, query issuing time, the number of intervening queries or picks, association order, and the number of sessions sharing such an association. The Q2Q correlation can assist in the generation of query suggestions in search engine result pages.

For example, "Electronic eavesdropping" is linked to the query "eavesdropping devices." "Hotel Mariot," spelled incorrectly, is related to "Hotel Marriott."

- Pick-to-pick (P2P). This correlation is similar to the Q2Q correlation in that it associates picks issued during a user session. For certain topics, the P2P correlation can reveal a group of URLs that are similar or related. For example, toyota.com is the top result when searching for "toyota car" , honda.com and ford.com may also be listed as similar results if users are looking for products that are similar.

Despite the fact that click-through data is often noisy and incomplete due to its sparse structure, it still provides useful information for improving search quality. Ask.com, which acquired DirectHit.com in 1999, was the first company to rank search results based on URL clickthrough rates. The adjustment is made when URLs with a lower rank receive more clicks than those with a higher rank. With Q2P and Q2Q correlations, Ask.com used session-based click data to improve ranking [21]. [22] investigated the use of click data to determine a user's relative preference in order to train a ranking function. In learning to rank, [23] looked at query reformulation that expands the Q2P correlations. [24] demonstrated the impact of integrating a number of click-based features on relevancy. [25, 26] investigated the bipartite graph structure of query-document relationships derived from a query log for modeling Q2P, Q2Q, and P2P correlations. [27] considered integrating the browsing activities of a search session for ranking purposes.

3.1.3.1.5 Rank Learning and Aggregation of Scores

In order to respond to a query, many ranking signals can be extracted or characterized for each matched document. These signals must be combined, and their weights must be determined either empirically or using a machine learning method. The signal aggregation can be done in a hierarchical manner. In this way, ranking signals can be fine-tuned by dividing them into several semantically-relevant components. Machine-learned ranking employs supervised or semi-supervised techniques to generate a ranking model automatically from training data. A set of queries makes up the training data, and each query has a list of documents with a relevance label. "Relevant", "related" or "irrelevant" for a three-level judgment are examples of relevance labels. There are several ways to model a learning function based on a training dataset, and multiple models can be combined using an ensemble method like bagging and boosting [28,29,30] to improve learning accuracy even further.

3.1.4 Retrieving Results

The final step in the operation of a search engine is to get the results, which is simply to display the results in a specific order in the browser. Search results pages sorted by search engines from the most relevant to the least relevant.

The advantage of crawlers are:

- They have a large number of pages.

- Simplicity of use.

- Familiarity. The majority of people who use the Internet to search are familiar with Google.

There are several disadvantages to crawlers:

- Sometimes, it's just too much information.

- Fooling the crawler is simple. Website hidden data can be manipulated to make the page appear to be something it is not. As a result, the Descartes search result may direct you to an undesired website.

- It is possible to change page rank. While search engine companies frown on the practice, there are ways to improve the position of your page in the list of results.

Examples of Crawler Based Search Engines:

- Google
- Bing
- Yahoo!
- Baidu
- Yandex

Aside from these well-known search engines, there are numerous crawler-based search engines available, such as DuckDuckGo, AOL, and Ask.

3.2 Human Powered Directories

Human powered directories, also known as open directory systems, rely on human-based activities to generate listings. The indexing in human powered directories works as follows:

- The site owner submits a brief description of the site to the directory, along with the category in which it will be listed.

- Sites that are submitted are manually reviewed and either added to the appropriate category or rejected for listing.

- Keywords entered into a search box will be matched with site descriptions. This means that changes made to the content of a web page are ignored because only the description is important.

- A good site with good content is more likely to get a free review than a bad site with bad content.

Human-powered directories such as Yahoo! Directory and DMOZ are prime examples. Google, unfortunately, has removed this method from their system.

The advantage are:

- Before being included, each page is reviewed for relevance and content. This means there will be no more unexpected sites.

- Having fewer results can sometimes mean finding what you're looking for faster.

There are several disadvantages:

- Delay in setting up a website and putting it in the directory.
- Unfamiliar design and format.

- May have difficulty with more confusing searches.

3.3 Hybrid Search Engines

Search engines use a combination of crawler-based and manual indexing to list web pages in search results. Crawler-based search engines, such as Google, rely on crawlers as their primary mechanism, with human-provided directories acting as a backup. Google, for example, may use human-powered directories to display a web page description in search results. Hybrid search engines are increasingly crawler-based as human-powered directories fade away.

On the other hand ,manual search results filtering is still in place to weed out duplicated and spam websites. When a website is flagged as spam, the owner must take action and resubmit it to search engines. Experts will manually review the submitted website before re-rendering it in the search results. While the crawler is in charge of the processes, the control over monitoring and displaying organic search results is left to the user.

3.4 Other Types of Search Engines

Aside from the three main categories mentioned above, search engines can be divided into a variety of groups based on how they are used. The following are some examples:

- To only show images, videos, news, products, and local listings, search engines use a variety of bots. The Google News page, for example, can only be used to find news from various newspapers.

- Some search engines, such as Dogpile, gather meta data from other search engines and directories in order to display it in search results. The name for this type of search engine is Metasearch.

- Semantic search engines like Swoogle provide accurate search results by deciphering the context of search queries.

3.5 Analyzing Search Engine Algorithms of some popular company

In this section, we will learn about how some popular search engine websites work.
Google Search Algorithm

Google is the world's most popular search engine. Their search engine has a market share of more than 90% on a regular basis. Every day, new websites are launched. Google can find these pages by following links from previously crawled content or by submitting a sitemap directly from the website owner. Any changes to existing content can also be sent to Google by requesting a recrawl of a specific URL. When the web crawlers have gathered enough data, they send it back to Google to be indexed.

The process of indexing begins with the examination of website data, which includes written content, images, videos, and the technical structure of the site. To figure out what any page they crawled is about, Google looks for positive and negative ranking signals like keywords and website freshness. Google's website index is made up of billions of pages and 100 million gigabytes of data. Google uses RankBrain, a machine-learning algorithm, and Knowledge Graph, a knowledge base, to organize this information. This all contributes to Google's ability to provide users with the most relevant content possible. They proceed to the ranking action after the indexing is complete.
Before a user interacts with Google's search functionality, everything that has happened up to this point has occurred in the background. The action that takes place based on what a user is looking for is known as ranking. When someone conducts a search, Google considers five main factors:

- **Query meaning**:Any end user's question's intent is determined in this way. When someone conducts a search, Google uses this information to figure out exactly what they want. Each query is parsed using sophisticated language models based on previous searches and usage patterns.

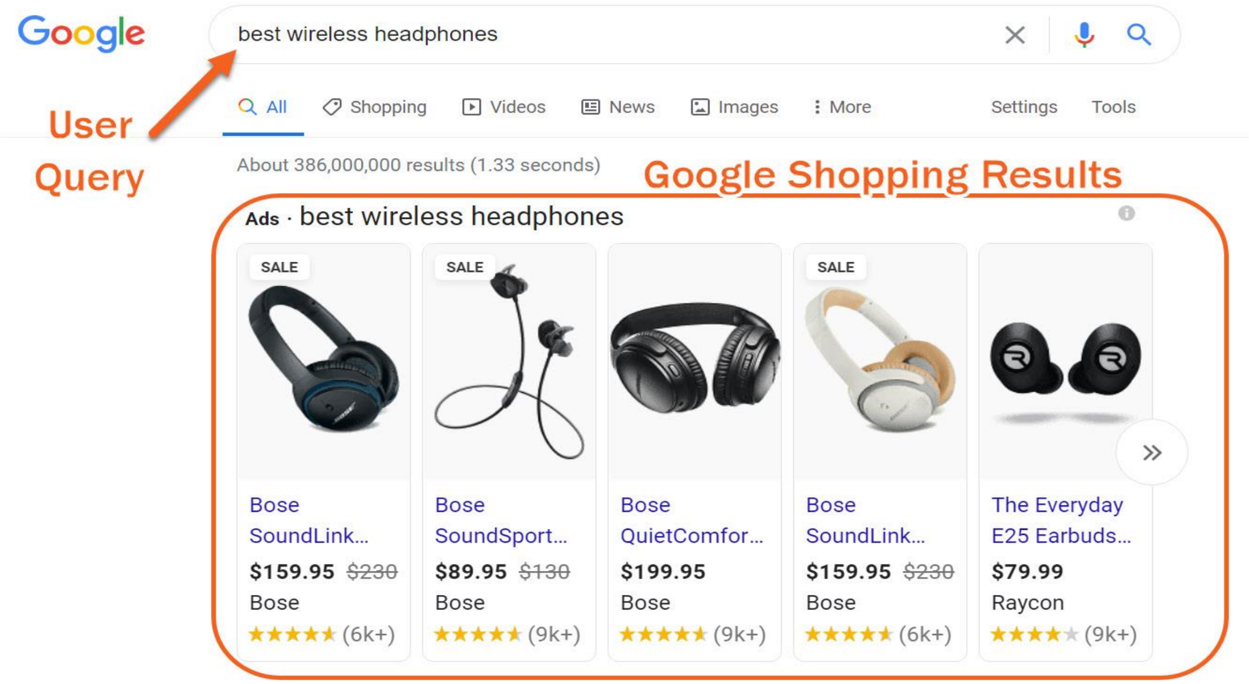
- **Web page relevance** When Google has figured out what a user's search query is about, they look at the content of ranking web pages to see which one is the most relevant. Keyword analysis is the driving force behind this. The keywords on a website must match Google's interpretation of the question posed by the user.

- **Content quality**: After matching keywords, Google goes one step further and assesses the content quality on the relevant web pages. This allows them to prioritize which results are displayed first by considering a website's authority, page rank, and freshness.

- **Web page usability**: Google prioritizes user-friendly websites in its search results. The term "usability" refers to everything from the speed of a website to its responsiveness.

- **Additional context and settings** This step customizes searches based on previous user interactions and Google platform preferences. Furthermore, in order to provide the best user experience and the most relevant answers, Google takes into account factors such as the user's location, language, and device (desktop or phone).

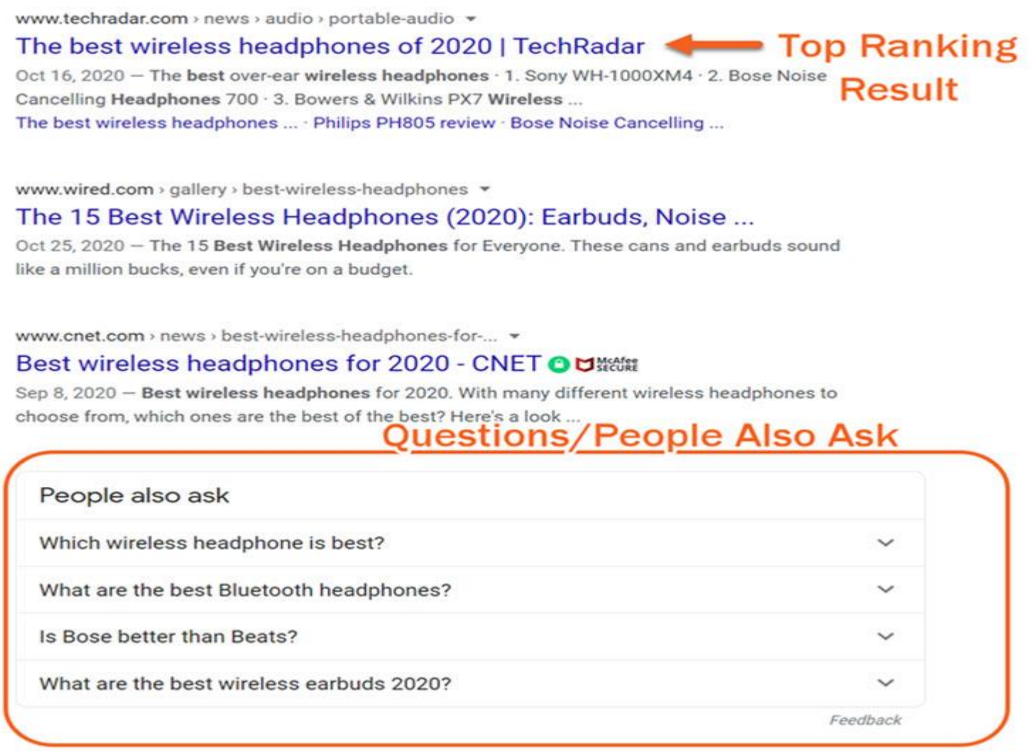
Google will return results that look something like this once all of this data has been processed:



best wireless headphones

About 386,000,000 results (1.33 seconds)

Figure 8:Search for "best wireless headphones" on Google. [31]



www.techradar.com › news › audio › portable-audio

The best wireless headphones of 2020 | TechRadar

Oct 16, 2020 — The **best** over-ear **wireless headphones** · 1. Sony WH-1000XM4 · 2. Bose Noise Cancelling **Headphones** 700 · 3. Bowers & Wilkins PX7 **Wireless** ...

The best wireless headphones ... · Philips PH805 review · Bose Noise Cancelling ...

Top Ranking Result

www.wired.com › gallery › best-wireless-headphones

The 15 Best Wireless Headphones (2020): Earbuds, Noise ...

Oct 25, 2020 — The 15 **Best Wireless Headphones** for Everyone. These cans and earbuds sound like a million bucks, even if you're on a budget.

www.cnet.com › news › best-wireless-headphones-for-...

Best wireless headphones for 2020 - CNET

Sep 8, 2020 — **Best wireless headphones** for 2020. With many different wireless headphones to choose from, which ones are the best of the best? Here's a look ...

Questions/People Also Ask

People also ask

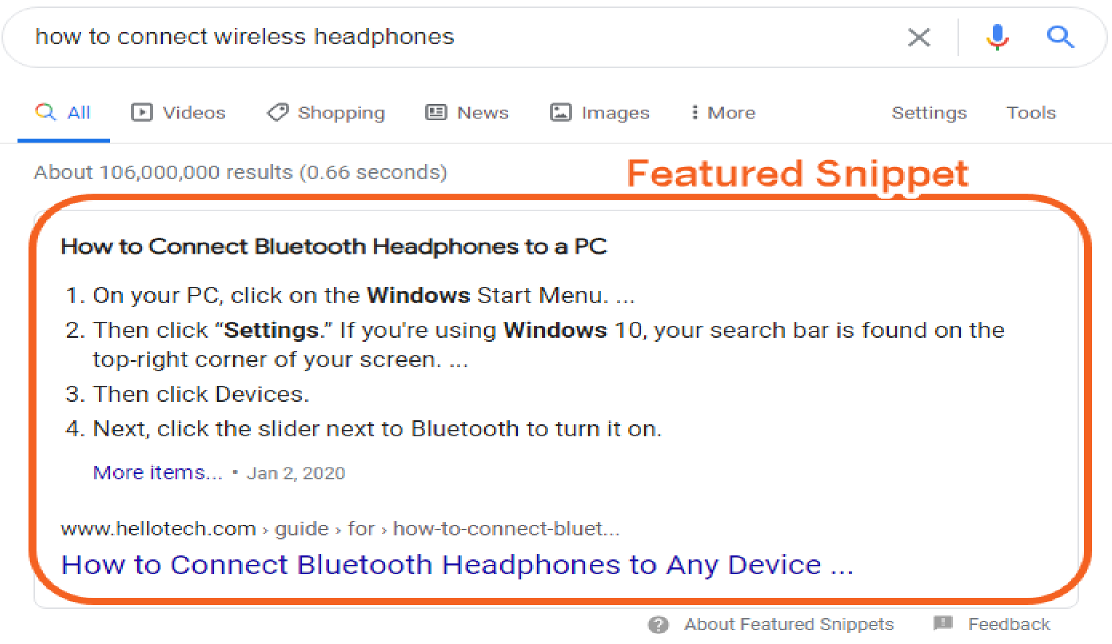
Which wireless headphone is best?

What are the best Bluetooth headphones?

Is Bose better than Beats?

What are the best wireless earbuds 2020?

Figure 9:Featured ranking and asking from " best wireless headphones " Google search. [31]



how to connect wireless headphones

About 106,000,000 results (0.66 seconds)

Featured Snippet

How to Connect Bluetooth Headphones to a PC

1. On your PC, click on the **Windows** Start Menu. ...
2. Then click "**Settings**." If you're using **Windows** 10, your search bar is found on the top-right corner of your screen. ...
3. Then click Devices.
4. Next, click the slider next to Bluetooth to turn it on.

More items... • Jan 2, 2020

www.hellotech.com › guide › for › how-to-connect-bluet...

How to Connect Bluetooth Headphones to Any Device ...

Figure 10:Featured snippet from "how to connect wireless headphones" Google search.[31]

Let's take a closer look at these findings:

- **Google query**: The question posed by a user.
- **Google shopping**: This query is interpreted by Google as someone looking to buy something. As a result, products matching this intent are pulled from their index and displayed first in the results.
- **Feature snippet**: Google displays specific information from a SERP result to allow users to review it without leaving the page.
- **Top-ranking results**: Google believes the first site listed in the results best matches the intent of a user's query. Based on the five ranking factors discussed earlier, the top-ranking result is the one that performs the best.

Because only Google has information on each of these pages in its index can these results be obtained. Google has analyzed websites to determine what keywords and intent they match for before a user performs a search. This process helps Google provide the most relevant content possible by quickly populating the results page after a search.

Bing Search Algorithm

Bing, Microsoft's proprietary search engine, surfaces results using Space Partition Tree And Graph (SPTAG), an open-source vector-search algorithm. Anyone can look at the nuts-and-bolts code that makes up Bing's search results and make comments because it's open source. The index builder and searcher modules are separated from the rest of the code:

- **Index Builder**: This is the code that organizes website data into vectors.
- **Searcher**: The method by which Bing connects search queries to vectors in their index.

Instead of using a keyword-first approach like Google, Bing stores and indexes information by breaking it down into individual data points called vectors. A vector is a numerical representation of a concept, and it is this concept that Bing's search structure is built on.

Bing's search queries are based on the Approximate Nearest Neighbor algorithm, which uses deep learning and natural-language models to provide faster results based on the proximity of certain vectors to one another.

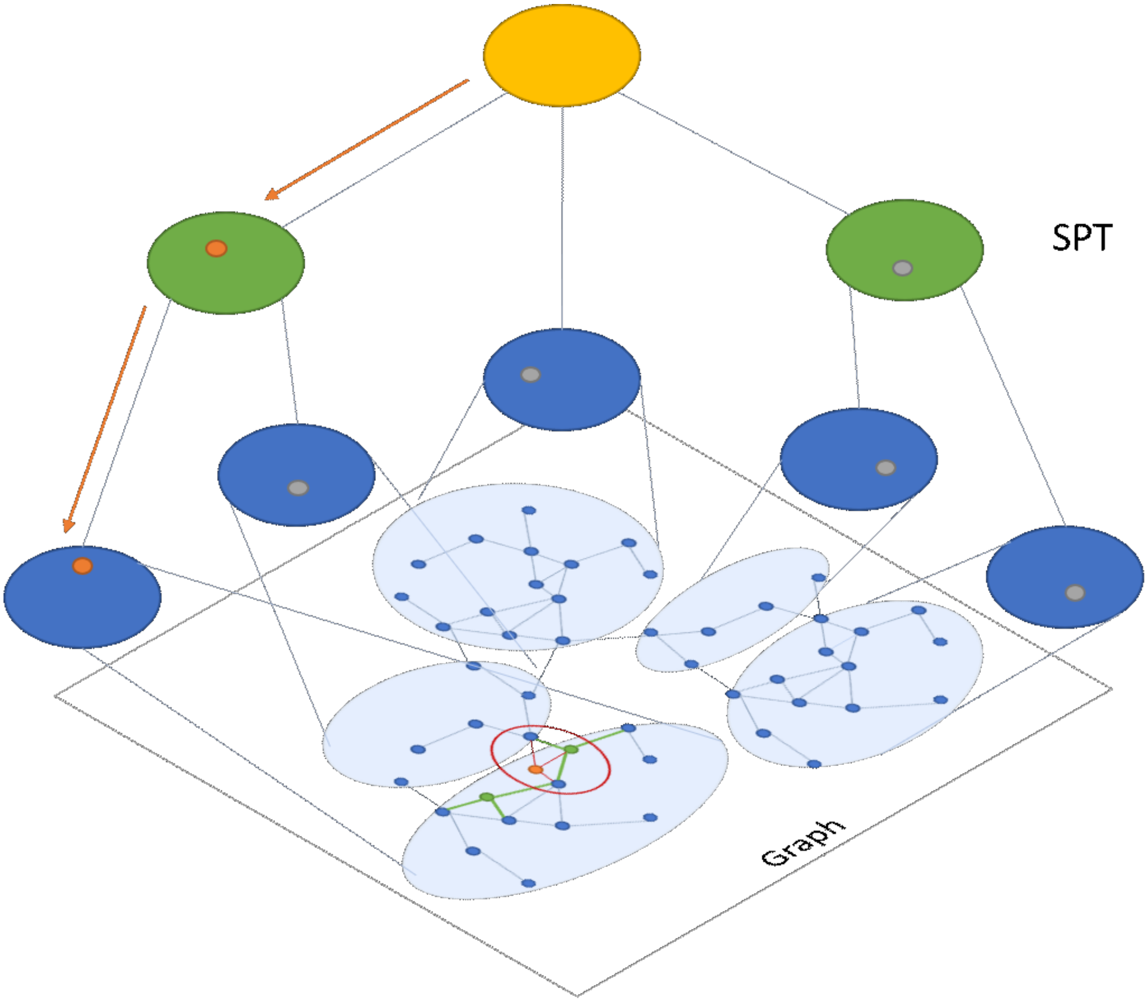
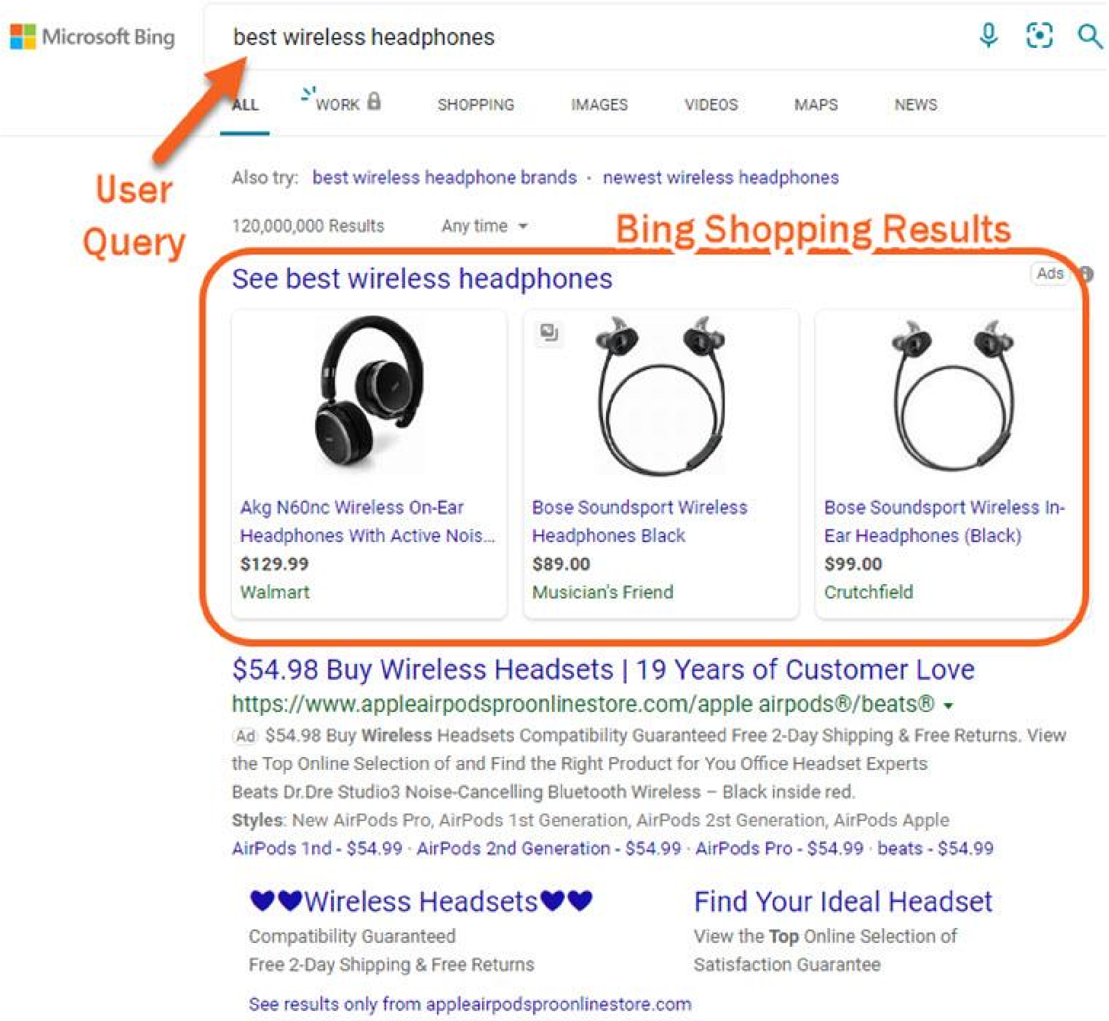


Figure 11:Featured snippet from "how to connect wireless headphones" Google search.[31]

When we consider the yellow dot as a user query, we can see that the green dots are the closest neighbors, followed by the blue dots. We can how Bing's algorithm decides which information is most relevant to the user's search by following the orange arrow.In generally, the process of building their database still follows the crawl, index, rank actions.

Bing crawls websites in order to find new or updated content. They then use vectors to store that data in their index. They examine specific ranking factors after that. The biggest difference between Bing and Google is that Bing does not include pages without ranking authority, which means that new pages with no backlinks to an existing page with more authority will have a harder time ranking.

When we run the same search on Bing, we get the following results:



best wireless headphones

Also try: best wireless headphone brands · newest wireless headphones

120,000,000 Results Any time

See best wireless headphones

Akg N60nc Wireless On-Ear Headphones With Active Nois...
$129.99
Walmart

Bose Soundsport Wireless Headphones Black
$89.00
Musician's Friend

Bose Soundsport Wireless In-Ear Headphones (Black)
$99.00
Crutchfield

$54.98 Buy Wireless Headsets | 19 Years of Customer Love

https://www.appleairpodsproonlinestore.com/apple airpods®/beats®

Ad $54.98 Buy **Wireless** Headsets Compatibility Guaranteed Free 2-Day Shipping & Free Returns. View the Top Online Selection of and Find the Right Product for You Office Headset Experts

Beats Dr.Dre Studio3 Noise-Cancelling Bluetooth Wireless – Black inside red.

Styles: New AirPods Pro, AirPods 1st Generation, AirPods 2st Generation, AirPods Apple

AirPods 1nd - $54.99 · AirPods 2nd Generation - $54.99 · AirPods Pro - $54.99 · beats - $54.99

♥♥Wireless Headsets♥♥

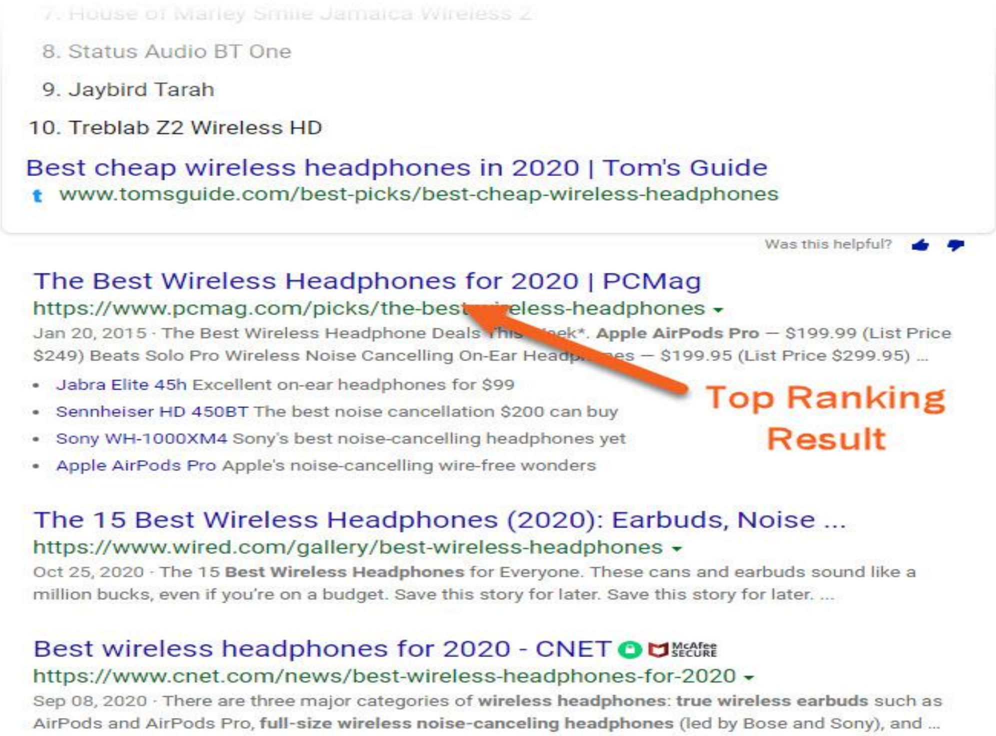
Compatibility Guaranteed
Free 2-Day Shipping & Free Returns

Find Your Ideal Headset

View the **Top** Online Selection of
Satisfaction Guarantee

See results only from appleairpodsproonlinestore.com

Figure 12:Search for "best wireless headphones" on Bing.[31]



7. House of Marley Smile Jamaica Wireless 2

8. Status Audio BT One

9. Jaybird Tarah

10. Treblab Z2 Wireless HD

Best cheap wireless headphones in 2020 | Tom's Guide
www.tomsguide.com/best-picks/best-cheap-wireless-headphones

The Best Wireless Headphones for 2020 | PCMag
https://www.pcmag.com/picks/the-best-wireless-headphones
Jan 20, 2015 · The Best Wireless Headphone Deals This Week*. **Apple AirPods Pro** — $199.99 (List Price $249) Beats Solo Pro Wireless Noise Cancelling On-Ear Headphones — $199.95 (List Price $299.95) ...

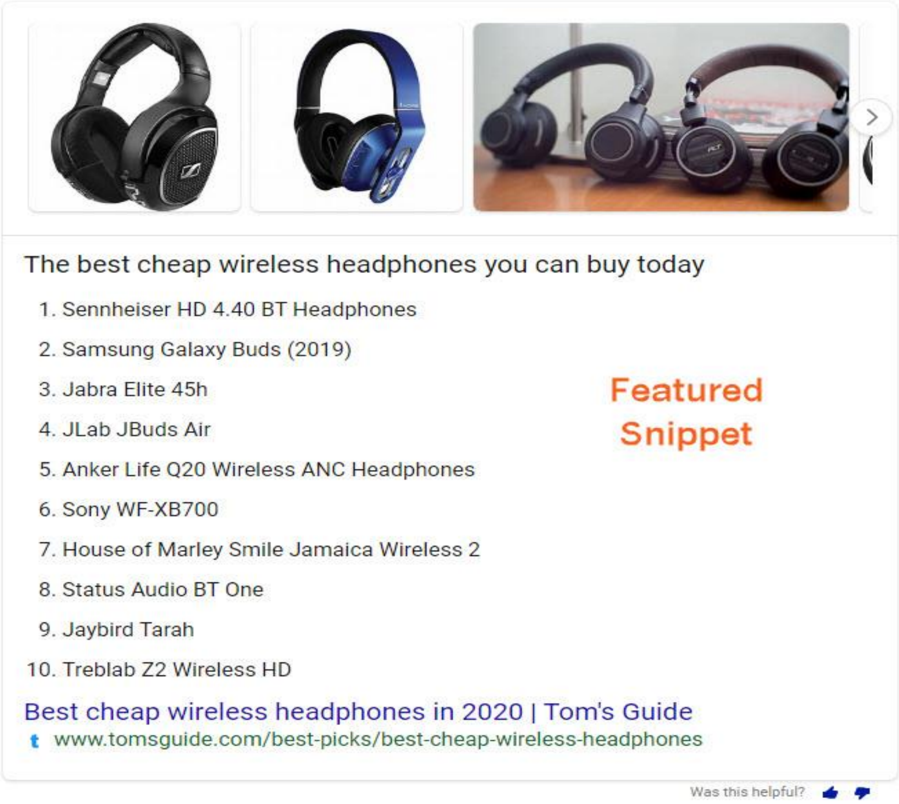
- Jabra Elite 45h Excellent on-ear headphones for $99
- Sennheiser HD 450BT The best noise cancellation $200 can buy
- Sony WH-1000XM4 Sony's best noise-cancelling headphones yet
- Apple AirPods Pro Apple's noise-cancelling wire-free wonders

Top Ranking Result

The 15 Best Wireless Headphones (2020): Earbuds, Noise ...
https://www.wired.com/gallery/best-wireless-headphones
Oct 25, 2020 · The 15 **Best Wireless Headphones** for Everyone. These cans and earbuds sound like a million bucks, even if you're on a budget. Save this story for later. Save this story for later. ...

Best wireless headphones for 2020 - CNET
https://www.cnet.com/news/best-wireless-headphones-for-2020
Sep 08, 2020 · There are three major categories of **wireless headphones**: **true wireless earbuds** such as AirPods and AirPods Pro, **full-size wireless noise-canceling headphones** (led by Bose and Sony), and ...

Figure 13: Featured ranking from " best wireless headphones " bing search. [31]



The best cheap wireless headphones you can buy today

1. Sennheiser HD 4.40 BT Headphones
2. Samsung Galaxy Buds (2019)
3. Jabra Elite 45h
4. JLab JBuds Air
5. Anker Life Q20 Wireless ANC Headphones
6. Sony WF-XB700
7. House of Marley Smile Jamaica Wireless 2
8. Status Audio BT One
9. Jaybird Tarah
10. Treblab Z2 Wireless HD

Featured Snippet

Best cheap wireless headphones in 2020 | Tom's Guide
www.tomsguide.com/best-picks/best-cheap-wireless-headphones

Figure 14:Featured snippet from " best wireless headphones " bing search.[31]

The structure of the results appears to be similar. Google search yielded different results.

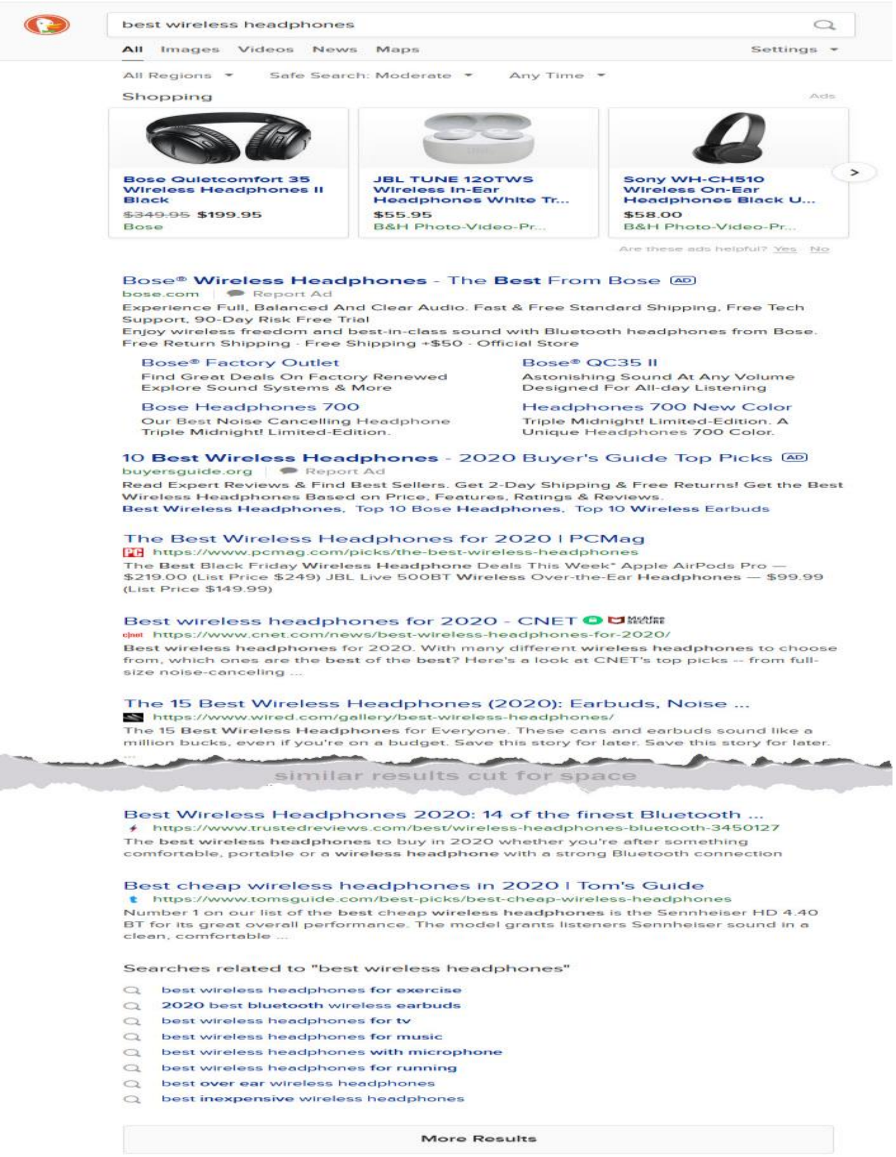
DuckDuckGo Search Algorithm

DuckDuckGo is the search engine to use if you're concerned about data privacy. While DuckDuckGo uses a proprietary web crawler called DuckDuckBot to scour web page content, much of the information shown on their results page comes from 400+ additional third-party sources like Bing, Yahoo, and Wikipedia.

DuckDuckGo, unlike Google and Bing, does not collect personal information about its users, such as search history or IP address.

DuckDuckGo can also be used for completely anonymous browsing using the Tor network or an onion service for even more privacy.

DuckDuckGo has the most streamlined results page so far as a result of this privacy focus:



Bose Quietcomfort 35 Wireless Headphones II Black
~~$349.95~~ $199.95
Bose

JBL TUNE 120TWS Wireless In-Ear Headphones White Tr...
$55.95
B&H Photo-Video-Pr...

Sony WH-CH510 Wireless On-Ear Headphones Black U...
$58.00
B&H Photo-Video-Pr...

Bose® Wireless Headphones - The Best From Bose AD
bose.com | Report Ad
Experience Full, Balanced And Clear Audio. Fast & Free Standard Shipping, Free Tech Support, 90-Day Risk Free Trial
Enjoy wireless freedom and best-in-class sound with Bluetooth headphones from Bose. Free Return Shipping · Free Shipping +$50 · Official Store

Bose® Factory Outlet
Find Great Deals On Factory Renewed Explore Sound Systems & More

Bose® QC35 II
Astonishing Sound At Any Volume Designed For All-day Listening

Bose Headphones 700
Our Best Noise Cancelling Headphone Triple Midnight! Limited-Edition.

Headphones 700 New Color
Triple Midnight! Limited-Edition. A Unique Headphones 700 Color.

10 Best Wireless Headphones - 2020 Buyer's Guide Top Picks AD
buyersguide.org | Report Ad
Read Expert Reviews & Find Best Sellers. Get 2-Day Shipping & Free Returns! Get the Best Wireless Headphones Based on Price, Features, Ratings & Reviews.
Best Wireless Headphones, Top 10 Bose Headphones, Top 10 Wireless Earbuds

The Best Wireless Headphones for 2020 | PCMag
https://www.pcmag.com/picks/the-best-wireless-headphones
The Best Black Friday Wireless Headphone Deals This Week* Apple AirPods Pro — $219.00 (List Price $249) JBL Live 500BT Wireless Over-the-Ear Headphones — $99.99 (List Price $149.99)

Best wireless headphones for 2020 - CNET
https://www.cnet.com/news/best-wireless-headphones-for-2020/
Best wireless headphones for 2020. With many different wireless headphones to choose from, which ones are the best of the best? Here's a look at CNET's top picks -- from full-size noise-canceling ...

The 15 Best Wireless Headphones (2020): Earbuds, Noise ...
https://www.wired.com/gallery/best-wireless-headphones/
The 15 Best Wireless Headphones for Everyone. These cans and earbuds sound like a million bucks, even if you're on a budget. Save this story for later. Save this story for later. ...

Best Wireless Headphones 2020: 14 of the finest Bluetooth ...
https://www.trustedreviews.com/best/wireless-headphones-bluetooth-3450127
The best wireless headphones to buy in 2020 whether you're after something comfortable, portable or a wireless headphone with a strong Bluetooth connection

Best cheap wireless headphones in 2020 | Tom's Guide
https://www.tomsguide.com/best-picks/best-cheap-wireless-headphones
Number 1 on our list of the best cheap wireless headphones is the Sennheiser HD 4.40 BT for its great overall performance. The model grants listeners Sennheiser sound in a clean, comfortable ...

Searches related to "best wireless headphones"

- best wireless headphones for exercise
- 2020 best bluetooth wireless earbuds
- best wireless headphones for tv
- best wireless headphones for music
- best wireless headphones with microphone
- best wireless headphones for running
- best over ear wireless headphones
- best inexpensive wireless headphones

Figure 15:Results from " best wireless headphones " DuckDuckGo search.[31]

Computational databases like WolframAlpha, a platform designed primarily to answer complex mathematical equations and provide data analysis tools, are among DuckDuckGo's 400 additional sources. Other sources include Instant Answers, which, like the feature snippets we've seen from Google and Bing, pull content from relevant websites in an attempt to provide on-page answers.
DuckDuckGo doesn't go into detail about the various ranking factors that go into these results pages, but it does suggest that linking to high-authority sites is a good idea.
Another intriguing feature of the DuckDuckGo platform is that it allows users to completely bypass the search results page by using custom parameters known as bangs. DuckDuckGo acts as a search portal for platforms such as Wikipedia, Amazon, and Twitter, by pulling information from multiple sources to display results. When you know you want to search on a different site, such as Wikipedia or Amazon, our bangs will get you there the quickest. If you search for !w filter bubble on Wikipedia, you'll find it there.

Chapter 4 Used technologies

4.1 Operating System

The operating system (OS) is the program that controls all of the computer's programs (software) and hardware:

- Hardware is CPU (central processing unit), memory (RAM - random access memory and ROM - read only memory), input devices (such as keyboard, mouse) and output devices (such as computer and laptop). print, speaker).
- Software is programs such as text editors (word, excel), web browsers (like chrome, firefox).

4.1.1 Windows

As we know ,Windows is a popular operating system for computers, supported and developed by Microsoft. It brings closeness and ease of use to the users. Therefore, I use Windows to execute my project.

4.2 Database

A database is a collection of data that has been organized for easy access, management, and updating. Computer databases typically store assemblages of data records or files containing information such as sales transactions, customer data, financials, and product information.
Databases are used to store, manage, and access all types of data. They amass data on people, places, and things. This data is gathered in one location so that it can be viewed and analyzed. Databases are a logically organized collection of data.

Database plays an extremely important role when working with data systems. They help the user to successfully connect the data. Users can access the database system faster and easier. Database is the source base for users to access the necessary information.
The main feature of the Database is to retrieve information and data by many different methods. The retrieved contents are guaranteed to have a high degree of data integrity. At the same time, the output information source is not duplicated at all, if any, the probability is also very low. The database allows multiple users to access concurrently at the same time.

4.2.1 SQL

SQL[32] is Structured Query Language, is a standard language for storing , modifying and retrieving data in databases. It is very popularly used in many current software such as MS SQL Server, Oracle, Sybase, MySQL, etc and is supported by most of today's programming languages. Example code shown below:

```
SELECT column_name
FROM tables_name
WHERE column_name LIKE pattern;
```

We use the SELECT command to get results from one or more columns in the database through the column_name parameter. Next, the FROM command will tell the database from which table the above columns are taken using the tables_name parameter. Then, to get the necessary results, the WHERE and LIKE commands will be conditional statements to help remove the results outside the above conditions.

4.2.2 MySQL Workbench

MySQL is an open-source relational database management system (RDBMS),using Structured Query Language (SQL).

MySQL is a high-speed, stable and easy-to-use database management system that is portable, works on many operating systems, provides a large system of very powerful functions.MySQL is also integrated with apache, PHP.

Let's learn in detail about the advantages of MySQL below:

- **Ease of Use**: MySQL is a high-speed, stable, easy-to-use and cross-operating database that provides a large set of very powerful utility functions.
- **High security**: MySQL is well suited for applications that have access to databases on the Internet when possessing many security features even at a high level.
- **Multi-feature**: MySQL supports many of the SQL functions expected from a relational database management system both directly and indirectly.
- **Scalable and powerful**: MySQL can handle a lot of data and moreover it can be extended if needed.
- **Fast**: The introduction of several standards allows MySQL to work very efficiently and cost-effectively, thus increasing execution speed.

Disadvantages of MySQL:

- **Limitations**: By design, MySQL is not intended to do it all and it comes with limitations on the functionality an application might need.
- **Reliability**: The way specific functions are handled with MySQL (e.g. references, transactions, audits, etc.) makes it less reliable than some DBMSs. other relationship.
- **Limited capacity**: If your number of records is growing, it is quite difficult to access your data, then we will have to apply many measures to speed up data retrieval such as load sharing this database to many servers, or create a MySQL cache.

4.3 Back-end

A server, an application, and a database make up the back end of a website. A back-end creates and maintains members as well as the technology that powers those components, allowing the user interface to function.

4.3.1 Visual Studio Code

Visual Studio Code is one of the most famous programming support tool was developed by Microsoft,is a source-code editor that can be used with a variety of programming languages, including Java, JavaScript, CSS, PHP, Node.js, Python and C++. It supports functional such as debugging, syntax highlighting, intelligent code completion, snippets, code refactoring and embedded Git.

4.3.2 PHP

PHP: Hypertext Preprocessor, is a popular scripting language that is used to develop website. Because it is optimized for web applications, is fast , flexible,easy to learn and practical so PHP powers can bring everythings to websites.

4.3.3 Laravel

Laravel is a free and open source PHP framework for Web Artisans, built to support the development of web softwares and applications, following the MVC(Model-View-Controller) software architecture. PHP framework is a library with resources available for each domain for programmers to use.

The advantages of Laravel are:

- Structure of frameworks: Laravel inherits the advantages and strengths of other Frameworks, when it has a very strong route section.
- Easy connect to Database: Supports connections across multiple database applications,such as MySQL,Postgres,SQLite and SQL server. We can use PHP code to control the Database, instead of using SQL.
- Rapid development: Provides many support features for programmers.
- High security : Use PDO to protect against SQL Injection attacks and use a hidden token field to prevent CSRF Request attacks.
- MVC : We will learn more about it in the MVC section.

4.3.4 MVC architecture

MVC (Model-View-Control) is often understood as a design pattern, requires the creation of three components,including:

- Model: has the function of storing all the data of the application and is the connection between the two components View and Controller. Specifically, Model contains all business logic, processing methods, database access, data description objects.
- View: displays the Model’s data to the user. Such as displaying a window, button or text. It includes anything that is visible to the user.

- Controller: exists between View and Model, responds to events (typically user-derived) and provides changes to Model. Based on these changes , View itself is updated.

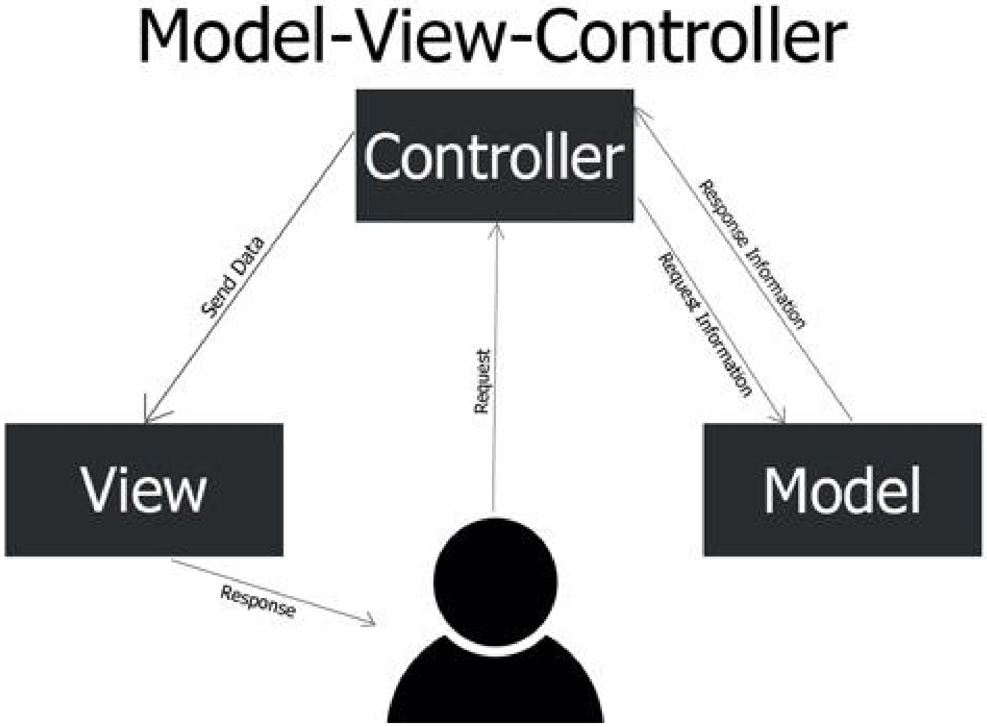


Figure 16:MVC pattern structure. [33]

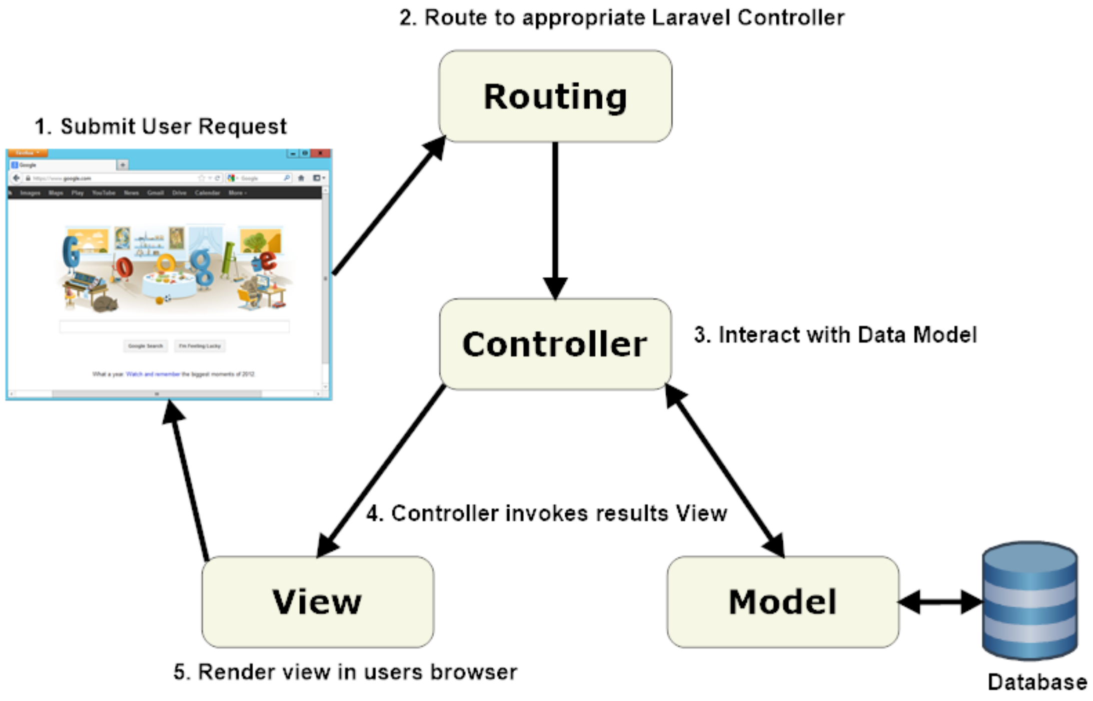

2. Route to appropriate Laravel Controller
Routing
1. Submit User Request
Controller
3. Interact with Data Model
4. Controller invokes results View
View
Model
5. Render view in users browser
Database

Figure 17:How MVC work[38]

4.4 Font-end

The part of a website that users interact with is called the front-end. Everything you see on the web, from fonts and colors to dropdown menus and sliders, is a combination of HTML, CSS, and JavaScript that is controlled by your computer's browser.

4.4.1 HTML and CSS

HTML:Hyper Text Markup Language, is known as the standard markup language for creating Web browser (as World Wide Web). Besides, It can be assisted by technologies such as CSS (Cascading Style Sheets) and JavaScript (Scripting Language).HTML describes the structure of a Web page,consists of a series of elements that tell the browser how to display the content. In addition, HTML also helps increase the ranking of the website thanks to search engine optimization.
CSS is the language we use to style an HTML document and describes how HTML elements should be displayed. Thanks to the combination between CSS and HTML , the content layout of the website becomes clearer, good-looking and easier to understand.

4.4.2 Jquery

jQuery is a JavaScript Library, simplifies JavaScript programming and easy to learn. jQuery makes browsing HTML documents easier. Simplifies a lot of the complicated things from JavaScript, like AJAX calls and DOM manipulation.

Chapter 5 Project Implementation

5.1 System analysis

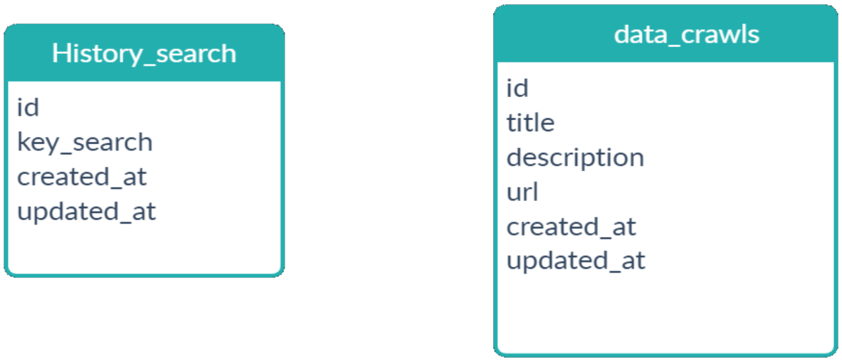
My system will be built as a website that uses search engines for product reviews. Therefore, the user interface design of the website will have simplicity as the main focus. Instead, focus on the main function that is the search engine and other functions that bring convenience and support the best user experience. Users can easily access the website through multiple browsers using devices such as computers. In general, the system will basically be built with two main roles including: administrator and user. Administrator will act as the manager, responsible for managing all systems including UI, back-end, font-end and database. The user will be the one who interacts with the website through the UI. Users will first visit the website, then enter the search bar with the product keyword they want. The site will return a searchable list of results. Each search result will be displayed in a shortened form that includes a title and description. If the user clicks on the title, the web page will redirect to the web page the user clicked on. On that website, users will see detailed product reviews. In addition, my website provides users with features such as supporting user experience such as popular search, results sorting or historical search.

The main idea of the website is to build a reputable product review website. This will bring a lot of benefits to users and sellers. For users, they will save time and easily find quality product reviews, so they can decide which product they should buy. In addition, this idea also has a huge impact on the areas of Marketing, Research, Sales or Product Promotion. For low-budget stores, sellers can decide what to trade on their own, saving costs and increasing income.

5.2 Database analysis

5.2.1 ERD - logical model

Figure 18 is a description of the structure of the ERD - logical model.



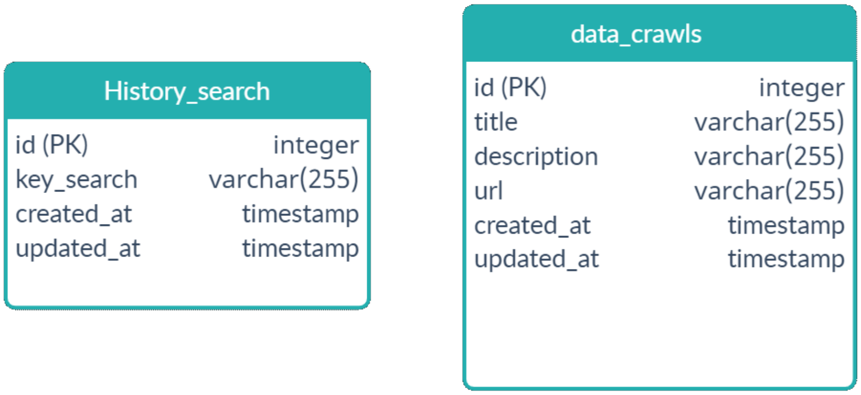
History_search
id
key_search
created_at
updated_at

data_crawls
id
title
description
url
created_at
updated_at

Figure 18: ERD - logical model

5.2.2 ERD - relational model

Figure 19 is a description of the structure of the ERD - relational model.



History_search	
id (PK)	integer
key_search	varchar(255)
created_at	timestamp
updated_at	timestamp

data_crawls	
id (PK)	integer
title	varchar(255)
description	varchar(255)
url	varchar(255)
created_at	timestamp
updated_at	timestamp

Figure 19: ERD - relational model

5.2.3 Data Dictionary

The table 5.1 display a key seach data:

Table 1:Key_search

Attribute	Data type	Length	Not Null	PK	FK	Description
Id	Interger	-	Yes	Yes	-	Represents the id of a history search
Key_search	Varchar	255	Yes	-	-	Represents input key search
Created_at	timestamp	-	-	-	-	Time of created
Update_at	timestamp	-	-	-	-	Time of updated

The table 5.2 display a data_crawl data:

Table 2: data_crawls

Attribute	Data type	Length	Not Null	PK	FK	Description
Id	Interger	-	Yes	Yes	-	Represents the id of a history search
Title	Varchar	255	Yes	-	-	Represents the title of web page
Url	Varchar	255	Yes	-	-	Represents the url of web page
Description	Varchar	255	Yes	-	-	Represents the desciption of web page
Created_at	timestamp	-	-	-	-	Time of created
Update_at	timestamp	-	-	-	-	Time of updated

5.3 Activity Diagram

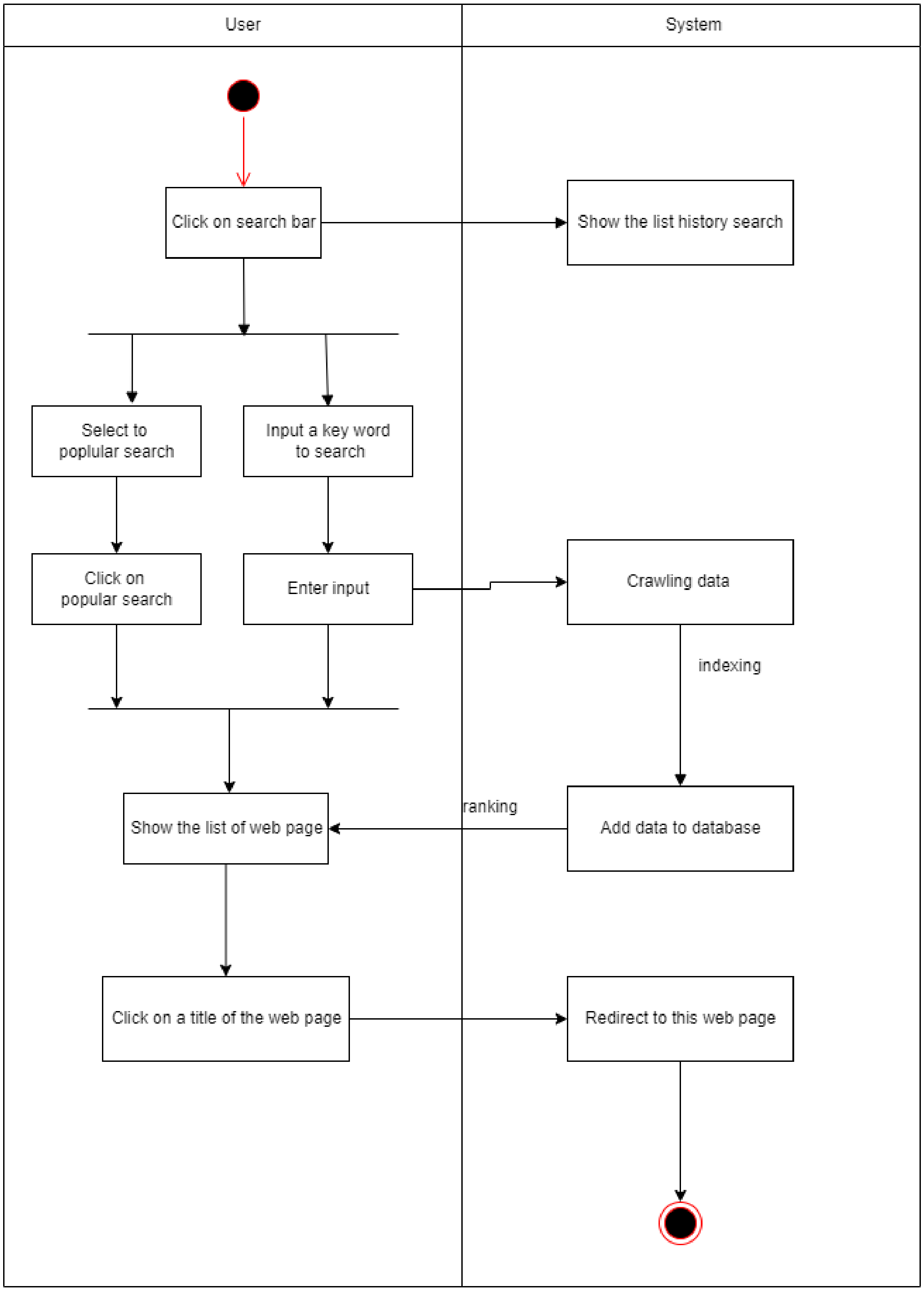
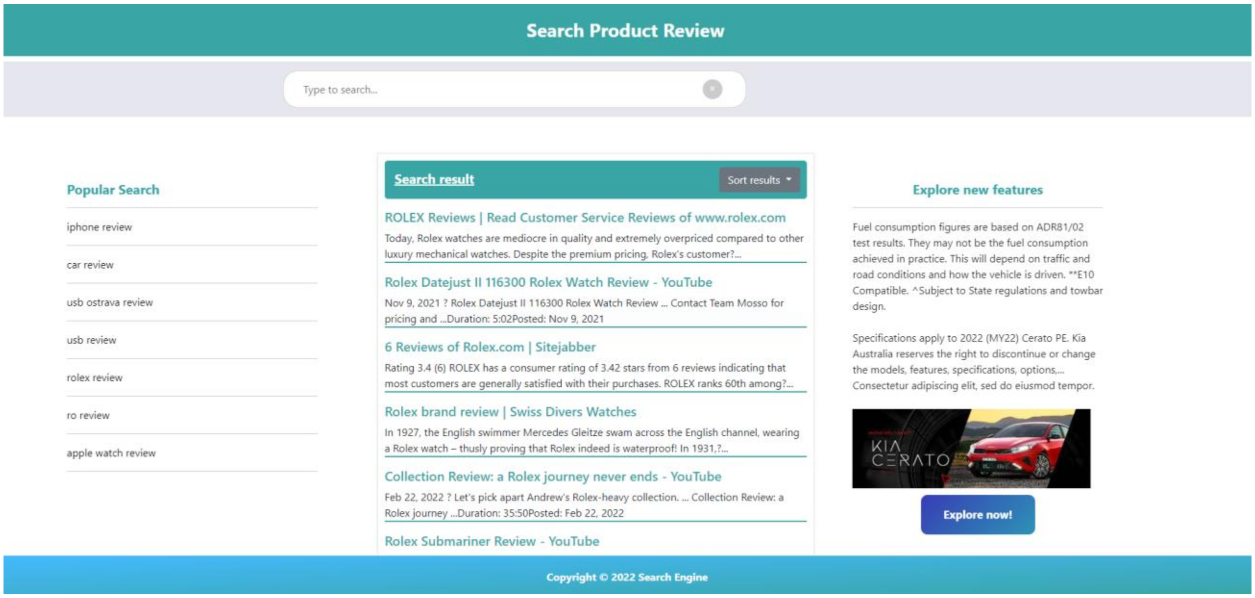

User
System
Click on search bar
Show the list history search
Select to poplular search
Input a key word to search
Click on popular search
Enter input
Crawling data
indexing
ranking
Show the list of web page
Add data to database
Click on a title of the web page
Redirect to this web page

Figure 20:Activity Diagram – Searching a web pag

5.4 Implement

On the website in the figure 21 below, we can easily see the functions and features clearly.



Search Product Review

Popular Search

iphone review

car review

usb ostrava review

usb review

rolex review

ro review

apple watch review

Search result

ROLEX Reviews | Read Customer Service Reviews of www.rolex.com

Today, Rolex watches are mediocre in quality and extremely overpriced compared to other luxury mechanical watches. Despite the premium pricing, Rolex's customer?...

Rolex Datejust II 116300 Rolex Watch Review - YouTube

Nov 9, 2021 ? Rolex Datejust II 116300 Rolex Watch Review ... Contact Team Mosso for pricing and ...Duration: 5:02Posted: Nov 9, 2021

6 Reviews of Rolex.com | Sitejabber

Rating 3.4 (6) ROLEX has a consumer rating of 3.42 stars from 6 reviews indicating that most customers are generally satisfied with their purchases. ROLEX ranks 60th among?...

Rolex brand review | Swiss Divers Watches

In 1927, the English swimmer Mercedes Gleitze swam across the English channel, wearing a Rolex watch – thusly proving that Rolex indeed is waterproof! In 1931,?...

Collection Review: a Rolex journey never ends - YouTube

Feb 22, 2022 ? Let's pick apart Andrew's Rolex-heavy collection. ... Collection Review: a Rolex journey ...Duration: 35:50Posted: Feb 22, 2022

Rolex Submariner Review - YouTube

Explore new features

Fuel consumption figures are based on ADR81/02 test results. They may not be the fuel consumption achieved in practice. This will depend on traffic and road conditions and how the vehicle is driven. **E10 Compatible. ^Subject to State regulations and towbar design.

Specifications apply to 2022 (MY22) Cerato PE. Kia Australia reserves the right to discontinue or change the models, features, specifications, options,... Consectetur adipiscing elit, sed do eiusmod tempor.

Copyright © 2022 Search Engine

Figure 21: Main page

5.4.1 Requirement

We need to install some of the following software to do the project:

- PHP runs on a WebServer environment and stores data through a database management system, so PHP usually comes with Apache, MySQL and Linux operating system (LAMP).
- On Windows, you can choose XamPP[34], specialized for programmers to install Apache-MySQL-PHP . You can see the installation instructions through there[35].

In my project, I use Visual Studio Code to do the project or you can refer to other Source code editors. We will perform the installation in the following steps:

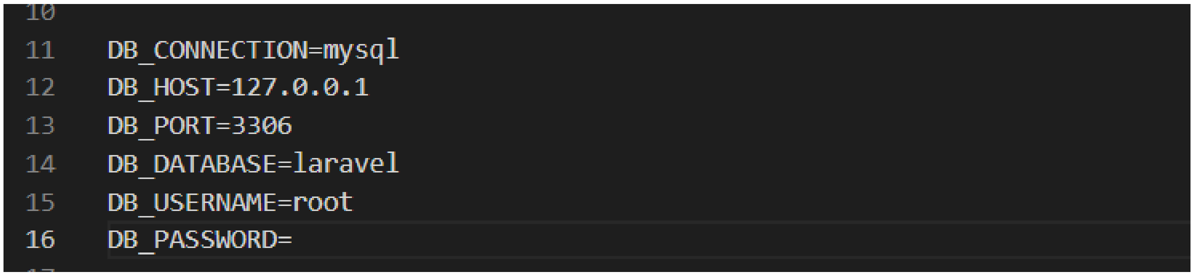
- Needing to install composer latest version[36]:Composer is a PHP Dependency Management tool that manages the libraries your Php project uses. More precisely Composer manages the dependencies of the resources in the project. It allows declaring the libraries that your project uses, composer will automatically load the code of the libraries. It generates the necessary files into your project, and updates the libraries when new versions are available.
- Download my project folder.
- In Visual Studio Code , we will open Command Prompt (CMD) and run the following commands to install laravel artisan :
  php artisan vendor:publish --tag=laravel-assets --ansi –force

php artisan key:generate --ansi

- Install the composer library into the project with the following command:
  composer require paquettg/php-html-parser --with-all-dependencies

- Start the server with the following command:
  php artisan serve

- Install MySQL and establish a connection to the database: First, we need to install MySQL[37]. In the project directory we have a .env file, this is the file that contains the environment settings, you will see a paragraph to set up the connection information to the database, change according to the project's information as follows:



```
DB_CONNECTION=mysql
DB_HOST=127.0.0.1
DB_PORT=3306
DB_DATABASE=laravel
DB_USERNAME=root
DB_PASSWORD=
```

Figure 22:Configure .env file

The settings include:

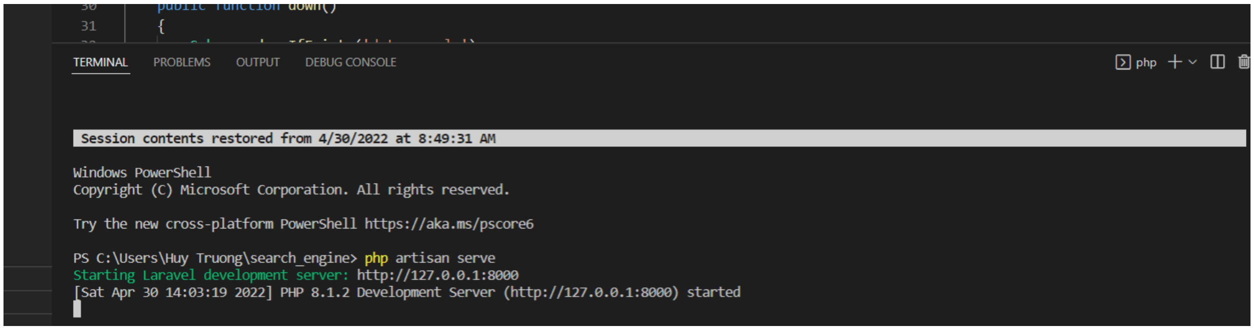
DB_CONNECTION: The type of database to connect to, can be MySQL, SQL Server, PostgreSQL, here we use MySQL.
DB_HOST: The IP address of the database server, we install MySQL on a personal computer, so the address 127.0.0.1 is the local IP address of our computer.
DB_PORT: The port of operation of the database, with MySQL defaulting to 3306.
DB_DATABASE: Database name.
DB_USERNAME: Database login name, with MySQL used on personal computers, use root user and white password.
DB_PASSWORD: Login Password

Then run the command to create database:

```
php artisan migrate
```

- Finally, we run the following command to start the server:

  php artisan serve



```
30      public function down()
31      {
```

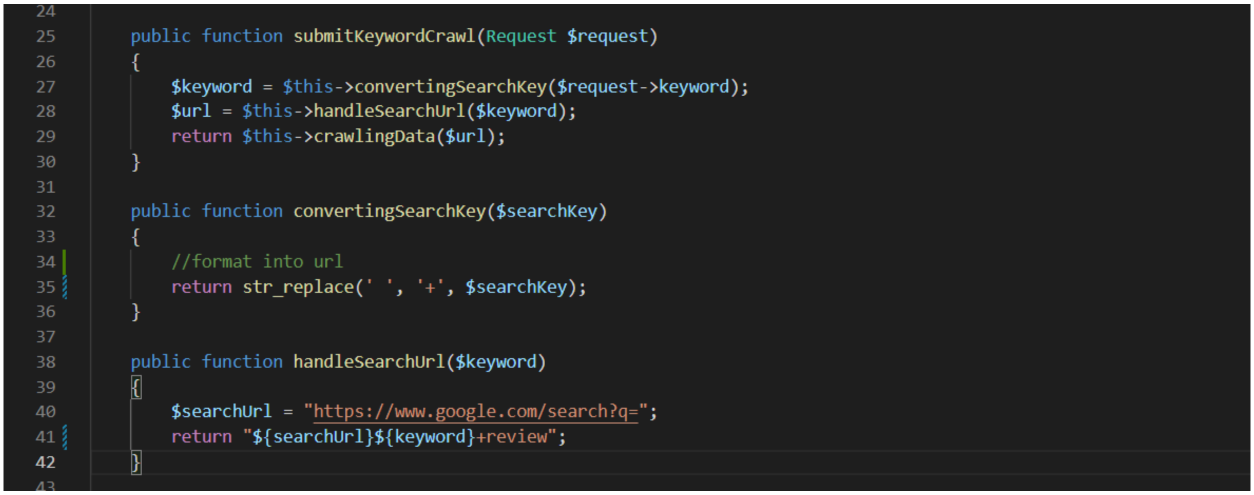
```
Session contents restored from 4/30/2022 at 8:49:31 AM

Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\Huy Truong\search_engine> php artisan serve
Starting Laravel development server: http://127.0.0.1:8000
[Sat Apr 30 14:03:19 2022] PHP 8.1.2 Development Server (http://127.0.0.1:8000) started
```

Figure 23:Start server on localhost

Now ,we can access to web page by localhost.

5.4.2 Handling url

Basically, the purpose of the split function is to get the user input after entering the search bar. After converting it to a complete url in Google's format, we will use that URL to crawl data from Google through the crawlingdata function.
In line 27 , it will call a convertSearchKey function passing the user input search as parameter, then format by removing spaces between characters to '+'. For example, 'Iphone 13 promax' convert to 'Iphone+13+promax'.
In line 28, it also calls a handleSearchUrl function to convert the user's keyword search into a complete url like 'https://www.google.com/search?q=iphone+13+promax+review'. Finally, we will call the crawlingdata function on line 29 to crawl data from the processed url itself.

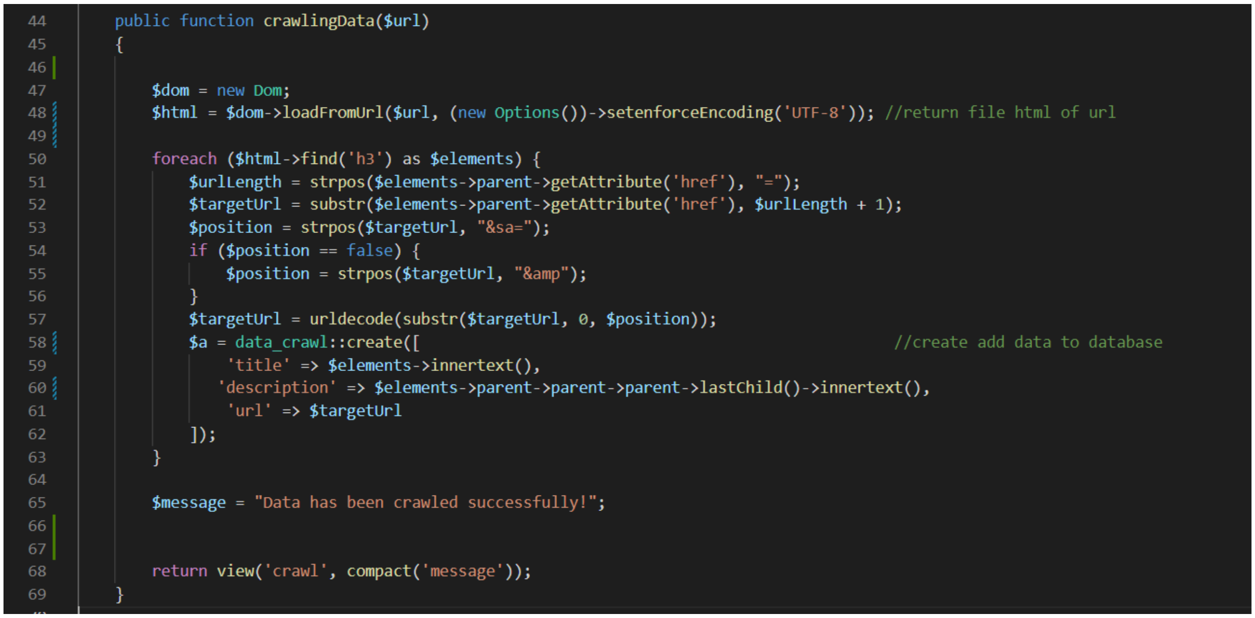


```
24
25      public function submitKeywordCrawl(Request $request)
26      {
27          $keyword = $this->convertingSearchKey($request->keyword);
28          $url = $this->handleSearchUrl($keyword);
29          return $this->crawlingData($url);
30      }
31
32      public function convertingSearchKey($searchKey)
33      {
34          //format into url
35          return str_replace(' ', '+', $searchKey);
36      }
37
38      public function handleSearchUrl($keyword)
39      {
40          $searchUrl = "https://www.google.com/search?q=";
41          return "${searchUrl}${keyword}+review";
42      }
43
```

Source code 1 :: Handling the searching of user

5.4.3 Crawling data

In this function, the main purpose is to get the necessary information such as title, description and url of each website in the list of results, then add this information to the database. I will explain the command lines and how they work in the steps below.
After we get the list of sites after crawling on line 48. From line 50 to line 57 we go to each page's tag then process the url string truncation step by step in order of the HTML element to get the required url. Finally, perform post-retrieval data addition that is stored in the database by columns like title, description, and url. In addition, the data after crawling, will be checked for duplicates or not. All of this will help reduce the data load.

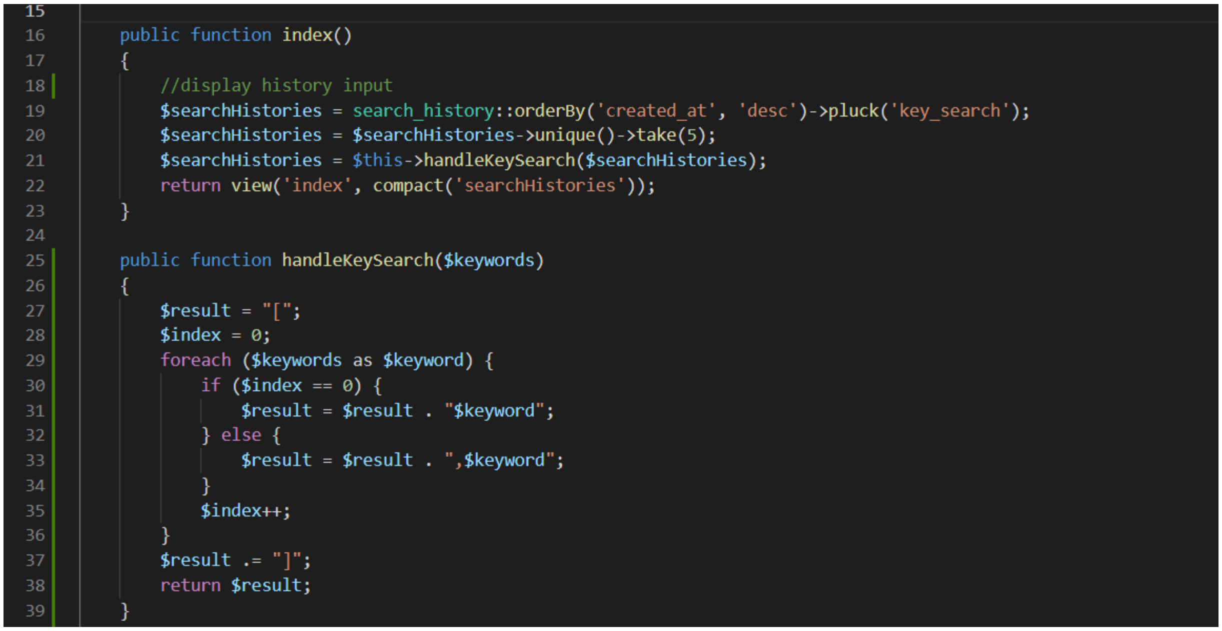


```
44      public function crawlingData($url)
45      {
46
47          $dom = new Dom;
48          $html = $dom->loadFromUrl($url, (new Options())->setenforceEncoding('UTF-8')); //return file html of url
49
50          foreach ($html->find('h3') as $elements) {
51              $urlLength = strpos($elements->parent->getAttribute('href'), "=");
52              $targetUrl = substr($elements->parent->getAttribute('href'), $urlLength + 1);
53              $position = strpos($targetUrl, "&sa=");
54              if ($position == false) {
55                  $position = strpos($targetUrl, "&amp");
56              }
57              $targetUrl = urldecode(substr($targetUrl, 0, $position));
58              $a = data_crawl::create([                                        //create add data to database
59                  'title' => $elements->innertext(),
60                 'description' => $elements->parent->parent->parent->lastChild()->innertext(),
61                  'url' => $targetUrl
62              ]);
63          }
64
65          $message = "Data has been crawled successfully!";
66
67
68          return view('crawl', compact('message'));
69      }
```

Source code 2:: Crawling data function

5.4.4 Displaying history search

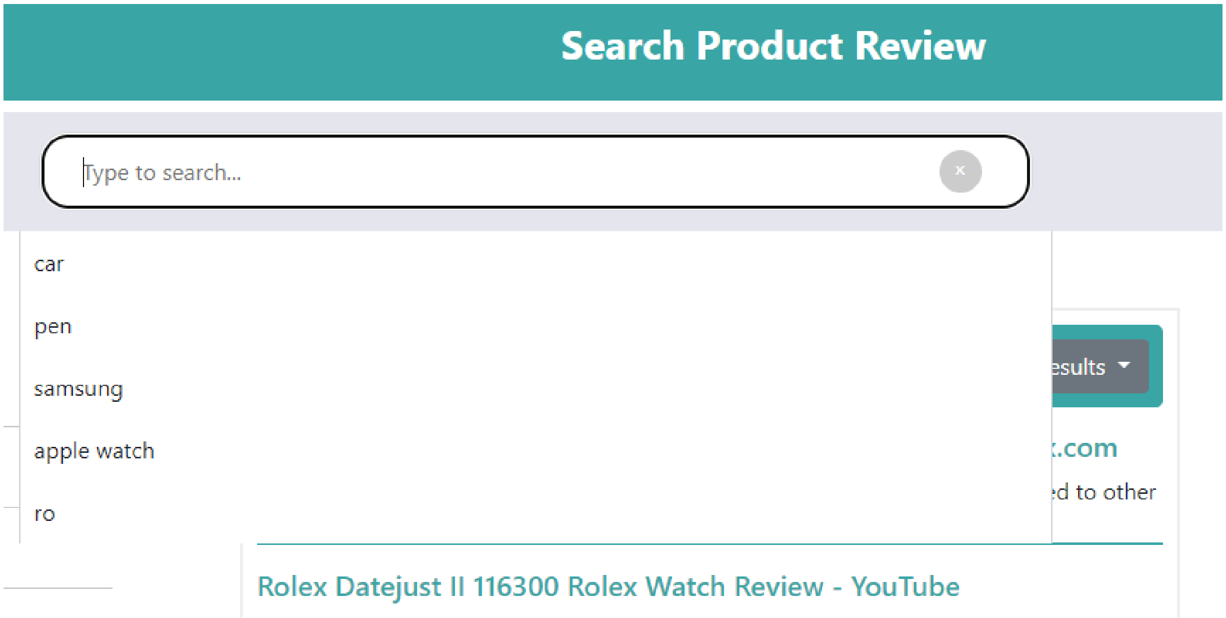
In this function, Its main effect will help users to review the keywords they have searched for before. Its main effect will help users to review the keywords they have searched for before.



```
15
16      public function index()
17      {
18          //display history input
19          $searchHistories = search_history::orderBy('created_at', 'desc')->pluck('key_search');
20          $searchHistories = $searchHistories->unique()->take(5);
21          $searchHistories = $this->handleKeySearch($searchHistories);
22          return view('index', compact('searchHistories'));
23      }
24
25      public function handleKeySearch($keywords)
26      {
27          $result = "[";
28          $index = 0;
29          foreach ($keywords as $keyword) {
30              if ($index == 0) {
31                  $result = $result . "$keyword";
32              } else {
33                  $result = $result . ",$keyword";
34              }
35              $index++;
36          }
37          $result .= "]";
38          return $result;
39      }
```

Source code 3: Displaying history search

In line 19 , the search_history table in database will perform the command to sort the data by the most recent created date. Then in line 20, we will get the top five results corresponding to the most recently created history search. Finally, the command on line 21 will be responsible for calling the handleKeysearch function to check the status of whether the user has entered the search bar or not. If not entered, the website is responsible for displaying the search history list to the user. The figure 24 below shows how history search works.



Search Product Review

Type to search...

car

pen

samsung

apple watch

ro

Rolex Datejust II 116300 Rolex Watch Review - YouTube

Figure 24:Start server on localhost

5.4.5 Boolean Full-Text Searches

With the IN BOOLEAN MODE modifier, MySQL can provide Boolean full-text searches. Certain characters at the beginning or end of a word in the search string have special meaning for this feature. In a query, operators indicate how a word should be searched.The boolean full-text search capability in the query supports the following operators: +,-,>,<,(),*, “” and no operator.
Meaning of operators in Boolean full-text search:

Table 3: Meaning of operators in Boolean full-text search

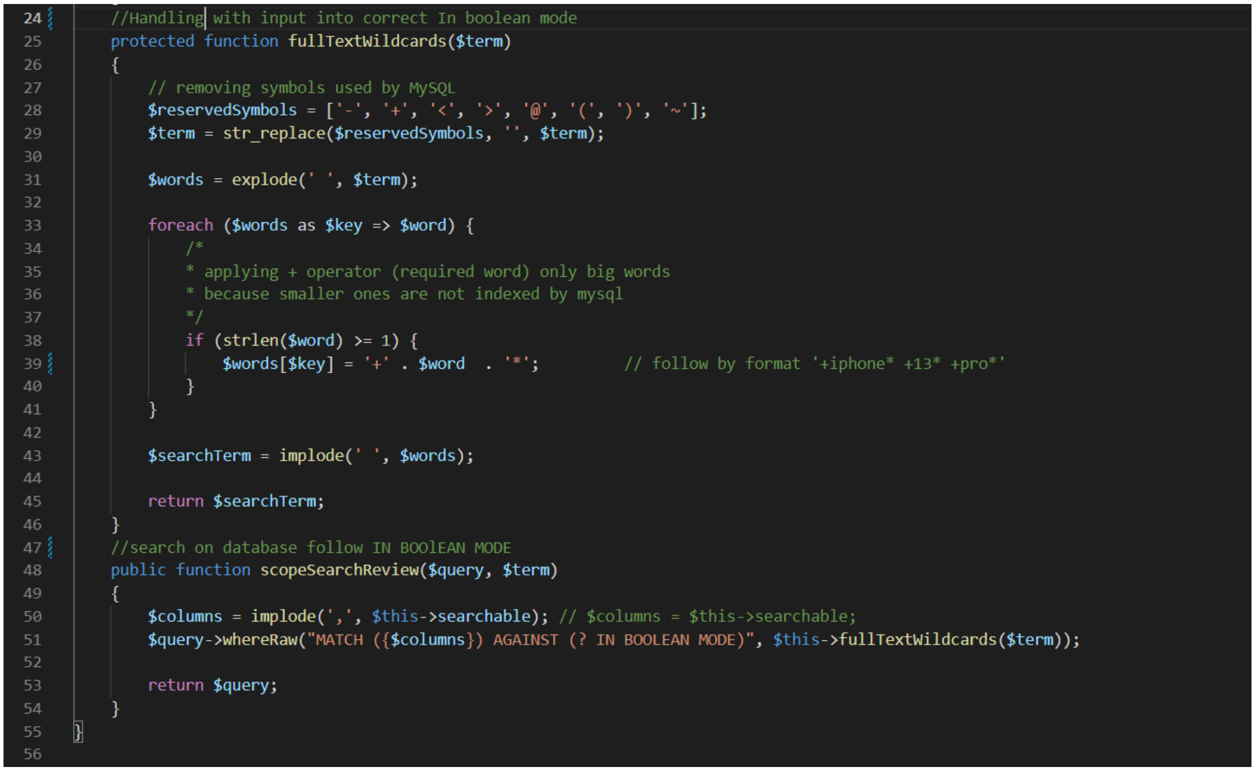
+	The word must appear.
-	The word must not appear.
> and <	> Including and increasing the rating for that word. And < Including and reducing the rating for that word.
()	Group words into subexpressions (allowing them to be included, excluded, ranked, etc. as a group).
~	Negation of a ranked word
*	Serves as a wildcard operator (or truncation). If words begin with the word preceding the * operator ,they will be matched.
“”	Identify a phrase (as opposed to a list of individual words, where the entire phrase is matched for inclusion or exclusion).
No operator	By default mode (do not use + or -) . The word can be placed optionally and the rows containing it are given a higher rank.

To search in my project, I combine + and *,for instance: ‘+iphone*’ . Means finding all rows that contain words such as ‘iphone’ , ‘iphoneX’,’iphone13’,...
How relevancy ranking is calculated: The inverse document frequency (IDF) value of a word and the term frequency (TF) value indicates how often a word occurs in a document.Therefore, the rows will be ranked based on the number of points. The higher the score, the higher the rank.

Using the following formula for single word: ${rank} = ${TF} * ${IDF} * ${IDF}
And ranking for multiple word search by the formula:
${rank} = (${TF} * ${IDF} * ${IDF}) + (${TF} * ${IDF} * ${IDF}) + (...) + ...

While ${IDF} = log10(${total_records} / ${matching_records}) ,where total_records is the number of records in the collection , matching_records is the number of records in which the search term occurs.

In the fullTextWildcards function , the input search will be formatted according to the + and * operator. Then, the scopeSearchReview function will be responsible for searching and retrieving data from the database.



```php
    //Handling with input into correct In boolean mode
    protected function fullTextWildcards($term)
    {
        // removing symbols used by MySQL
        $reservedSymbols = ['-', '+', '<', '>', '@', '(', ')', '~'];
        $term = str_replace($reservedSymbols, '', $term);

        $words = explode(' ', $term);

        foreach ($words as $key => $word) {
            /*
            * applying + operator (required word) only big words
            * because smaller ones are not indexed by mysql
            */
            if (strlen($word) >= 1) {
                $words[$key] = '+' . $word  . '*';         // follow by format '+iphone* +13* +pro*'
            }
        }

        $searchTerm = implode(' ', $words);

        return $searchTerm;
    }
    //search on database follow IN BOOlEAN MODE
    public function scopeSearchReview($query, $term)
    {
        $columns = implode(',', $this->searchable); // $columns = $this->searchable;
        $query->whereRaw("MATCH ({$columns}) AGAINST (? IN BOOLEAN MODE)", $this->fullTextWildcards($term));

        return $query;
    }
}
```

Source code 4: Searching algorithm

In addition , my project also has some small functions such as sorting alphabetically or a list of popular search . These functions all run on top of the above main functions and they are all executed in the background.

With the popular search function, I will show a list of the most searched words from users.

As for sorting by alphabet, I simply sort the list of results in order. I find this function quite interesting because it is very suitable for ecommercial search sites.

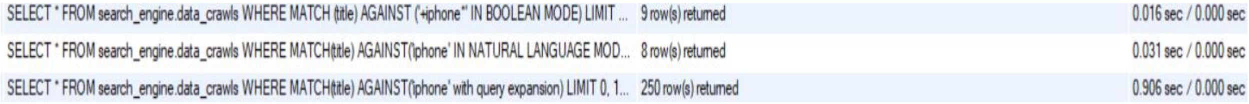
5.4.6 Experimental verification of algorithms and their comparison.

There are three types of full-text searches :

- Natural Language Full-Text Searches: If a key appears more than once in a record, the weight score of that keyword will increase, and if the keyword appears in many records, the weight score will be decreased.
- Boolean Full-Text searches: Search by search keyword. Use math to increase accuracy to decide which words will return results. It is possible to search for acronyms.
- Query expansion searches: Perform a search twice, in the second search MySQL will find a combination of the original search term with appropriate words that stand out from the original keyword. Find synonyms and acronyms. Correct spelling and re-type weight.

Each of those search techniques has its own set of caveats, and each of them may be better suited for different purposes. When deciding whether to use full-text searching, keep in mind that this type of searching has many subtleties of its own; be aware of both the benefits and drawbacks of using full-text searching in MySQL, and make an informed decision.

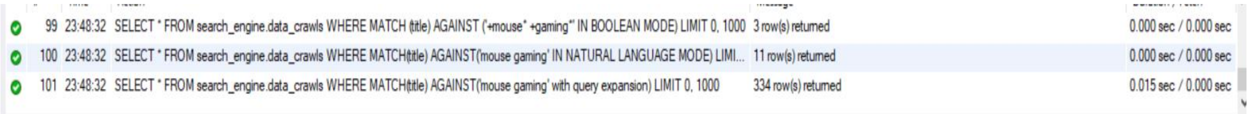
I will give a small comparison experiment below:
First, I will compare the search time and performance with a word in the same database.For example, search ‘iphone’ on figure below:



```
SELECT * FROM search_engine.data_crawls WHERE MATCH (title) AGAINST ('+iphone*' IN BOOLEAN MODE) LIMIT ...   9 row(s) returned     0.016 sec / 0.000 sec
SELECT * FROM search_engine.data_crawls WHERE MATCH(title) AGAINST('iphone' IN NATURAL LANGUAGE MOD...   8 row(s) returned     0.031 sec / 0.000 sec
SELECT * FROM search_engine.data_crawls WHERE MATCH(title) AGAINST('iphone' with query expansion) LIMIT 0, 1...   250 row(s) returned   0.906 sec / 0.000 sec
```

Figure 25:Results of search one word

The results are shown in Figure 25, showing that both Boolean's efficiency and time are superior to Natural's. As for Query Expansion, there are a lot of results, but the accuracy is not high because there are many results that do not contain the word 'iphone' and the search time is longer.



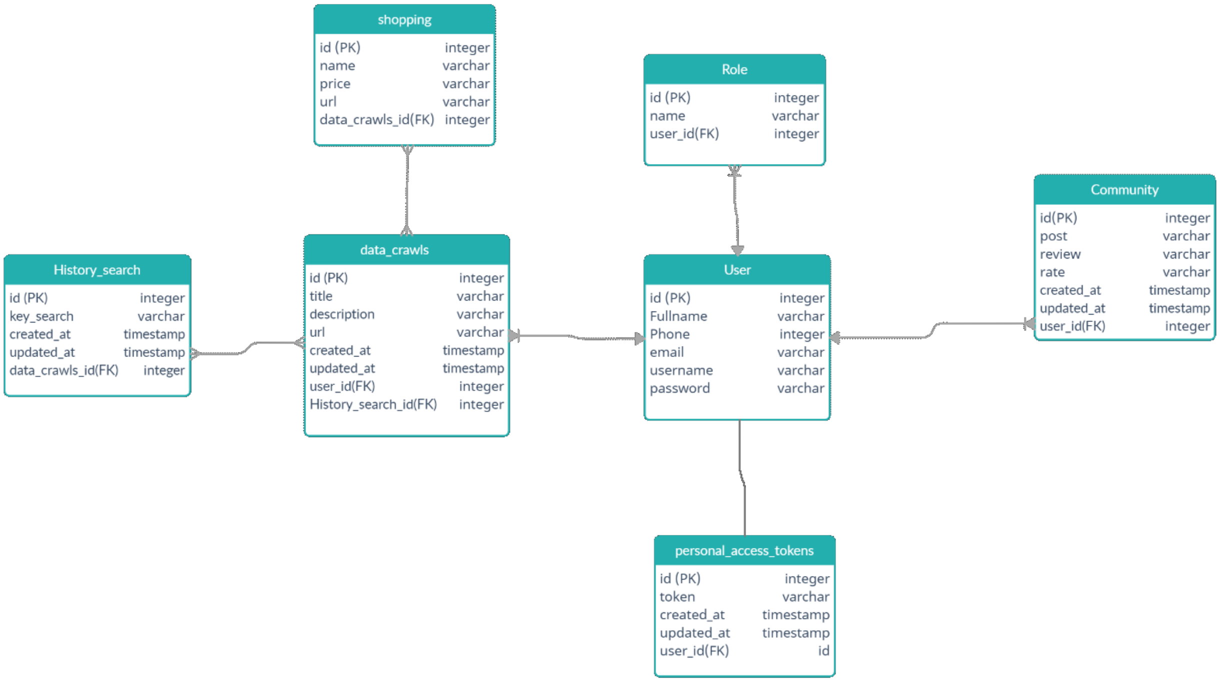
```
 99  23:48:32  SELECT * FROM search_engine.data_crawls WHERE MATCH (title) AGAINST ('+mouse* +gaming*' IN BOOLEAN MODE) LIMIT 0, 1000   3 row(s) returned     0.000 sec / 0.000 sec
100  23:48:32  SELECT * FROM search_engine.data_crawls WHERE MATCH(title) AGAINST('mouse gaming' IN NATURAL LANGUAGE MODE) LIMI...     11 row(s) returned    0.000 sec / 0.000 sec
101  23:48:32  SELECT * FROM search_engine.data_crawls WHERE MATCH(title) AGAINST('mouse gaming' with query expansion) LIMIT 0, 1000    334 row(s) returned   0.015 sec / 0.000 sec
```

Figure 26: Result of search two word

The results are shown in Figure 26, The search speed of Boolean and Natural are equal, but in terms of accuracy, Boolean is still superior.
Therefore, I find Boolean full-text search to be the most preferred method for my product currently.

Chapter 6 Discussion

In the future, my website can expand with more functions, providing users with more quality and reputable the product information. Besides, the user interface will also be improved to give users the best experience.
Specifically, the system will always update the algorithms to bring users more reliable and necessary information. In addition, the project will further develop product-related aspects such as open community or shopping that require users to have a personal account. In more detail, users can create their own articles to review and evaluate products according to their personal feelings. Shopping, on the other hand, will be where users can find products including prices and rates on sale from various websites. To do the above, the database system and user interface need to be upgraded.



shopping	
id (PK)	integer
name	varchar
price	varchar
url	varchar
data_crawls_id(FK)	integer

Role	
id (PK)	integer
name	varchar
user_id(FK)	integer

History_search	
id (PK)	integer
key_search	varchar
created_at	timestamp
updated_at	timestamp
data_crawls_id(FK)	integer

data_crawls	
id (PK)	integer
title	varchar
description	varchar
url	varchar
created_at	timestamp
updated_at	timestamp
user_id(FK)	integer
History_search_id(FK)	integer

User	
id (PK)	integer
Fullname	varchar
Phone	integer
email	varchar
username	varchar
password	varchar

Community	
id(PK)	integer
post	varchar
review	varchar
rate	varchar
created_at	timestamp
updated_at	timestamp
user_id(FK)	integer

personal_access_tokens	
id (PK)	integer
token	varchar
created_at	timestamp
updated_at	timestamp
user_id(FK)	id

Figure 27: New database

Chapter 7 Conclusion

In general, in this bachelor's thesis, I have analyzed, processed, designed, built and implemented about the Search Engine. In the analysis, I found out the purpose of the Search Engine built, it brings many benefits from many aspects such as entertainment, research or business to serve people's lives easily. Thanks to that I can see the importance and development potential of this engine. As for the design, the interface is built on subtle simplicity but gives users full functionality typical of Search Engine website. The combination of functions will bring convenience as well as usefulness to users. The logic of the system is built quite detailed and connected. In addition , we need to mention the important role of system analysis and construction . The analysis will help us understand in detail about the structure and operation of Search Engine, so that we can easily choose the right technology. This not only brings system optimization but also saves cost and time. More than that, analyzing the database helps me better understand what data is needed and estimate how much capacity the system can handle. Make sure there is no conflict between the system and the database. In general, I firmly believe that to be able to build an information search system is extremely complex, it requires programmers to have knowledge in many fields. Currently, most popular information search websites have a lot of different functions working together. Therefore, in order for these functions to work smoothly and correctly, developers need to understand in detail how they work and how they are linked. In fact, these systems need to be repaired and improved regularly to suit the needs of the user. To be honest, my system probably has a lot of bugs that have not been fixed so I look forward to improving and developing it more fully and popularly in the future.

As I mentioned above, my system needs a lot of time as well as resources to be able to develop it. Although, currently my website search is mainly focused on product reviews but I am sure it will be a good idea. Because it will bring a lot of influence to aspects such as business and research. In the future, I will develop the system to add an open community feature to discuss the product. As a result, consumers and merchants will have more choices for products. However, users will need to login and posts will be strictly managed by the admin. Besides, the shopping feature with product filtering function will also be added to bring users more convenience and usefulness. Therefore, I am confident that my Search Engine website will be a potential product and influence in the commercial environment.

Bibliography

[1] MySQL :: MySQL 5.7 reference manual :: 12.10 full-text search functions. (n.d.). Mysql.Com. Retrieved April 28, 2022, from https://dev.mysql.com/doc/refman/5.7/en/fulltext-search.html

[2] Wikipedia contributors. (2022, April 14). Main Page. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org//w/index.php?title=Main_Page&oldid=1082764418

[3] Nam L. (2020, April 22). SEO. https://vietmoz.edu.vn/seo-la-gi-va-tam-quan-trong-cua-seo/

[4] Bruce Croft, Donald Metzler, Trevor Strohman. Search Engines: Information Retrieval in Practice 1st Edition. ISBN-13: 978-0136072249.

[5] Search Engines. (n.d.). Tutorialspoint.Com. Retrieved April 28, 2022, from https://www.tutorialspoint.com/internet_technologies/search_engines.htm

[6] What is Search Engine?. (2018, March 2). Mona.solutions. https://mona.solutions/search-engine-la-gi-10-search-engine-pho-bien-nhat-gioi/

[7] Editorial Staff. (2014, May 23). What are Different Types of Search Engines? WebNots. https://www.webnots.com/what-are-different-types-of-search-engines/

[8] Sheppard, A. (2010). LibGuides: Searching the Internet: Types of search engines. https://libguides.astate.edu/c.php?g=14516&p=78177

[9] Yang, T. (n.d.). Web search engines: Practice and experience. Ucsb.Edu. Retrieved April 28, 2022, from https://sites.cs.ucsb.edu/~tyang/papers/2013bookchapter.pdf

[10] Jayant Madhavan, David Ko, Lucja Kot, Vignesh Ganapathy, Alex Rasmussen, and Alon Y. Halevy. Google's deep web crawl. PVLDB, 1(2):1241–1252, 2008.

[11] Sitemaps. http://www.sitemaps.org.

[12] robots.txt. http://www.robotstxt.org/.

[13] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. Inf. Process. Manage., 36(6):779–808, November 2000.

[14] Sepandar Kamvar, Taher Haveliwala, Chris Manning, and Gene Golub. Extrapolation methods for accelerating pagerank computations. In Twelfth International World Wide Web Conference (WWW 2003), 2003.

[15] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub. Exploiting the block structure of the web for computing pagerank. In Stanford University Technical Report, 2003.

[16] Joh M. Kleinberg. Authoritative sources in a hyperlinked environment. In Journal of the ACM, pages 604–632, 1999.

[17] Krishna Bharat and Monika R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98, pages 104–111, 1998.

[18] Brian D. Davison, Apostolos Gerasoulis, Konstantinos Kleisouris, Yingfang Lu, Hyun ju Seo, Wei Wang, and Baohua Wu. Discoweb: Applying link analysis to web search. In Proceedings of the Eighth International World Wide Web Conference, pages 148–149, 1999.

[19] Apostolos Gerasoulis, Wei Wang, and Hyun-Ju Seo. Retrieval and display of data objects using a cross-group ranking metric. US Patent 7024404, 2006.

[20] Tao Yang, Wei Wang, and Apostolos Gerasoulis. Relevancy-based database retrieval and display techniques. US Patent 7028026, 2006.

[21] Andy Curtis, Alan Levin, and Apostolos Gerasoulis. Methods and systems for providing a response to a query. US Patent 7152061, 2006.

[22] Thorsten Joachims. Optimizing search engines using clickthrough data. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM.

[23] Filip Radlinski and Thorsten Joachims. Query chains: learning to rank from implicit feedback. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, KDD '05, pages 239–248, New York, NY, USA, 2005. ACM.

[24] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. Learning user interaction models for predicting web search result preferences. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '06, pages 3–10, New York, NY, USA, 2006. ACM.

[25] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi, and WeiGuo Fan. Optimizing web search using web click-through data. In Proceedings of the thirteenth ACM international conference on Information and knowledge management, CIKM '04, pages 118–126, New York, NY, USA, 2004. ACM.

[26] Nick Craswell and Martin Szummer. Random walks on the click graph. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '07, pages 239–246, New York, NY, USA, 2007. ACM.

[27] Adish Singla, Ryen White, and Jeff Huang. Studying trailfinding algorithms for enhanced web search. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, pages 443–450, New York, NY, USA, 2010. ACM.

[28] Leo Breiman and E. Schapire. Random forests. In Machine Learning, pages 5–32, 2001.

[29] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. J. Mach. Learn. Res., 4:933–969, December 2003.

[30] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29:1189–1232, 2000.

[31] Condron, S. (2021, May 17). How search engines work: A guide to Google and Bing algorithms. The Official SpyFu Blog. https://www.spyfu.com/blog/how-do-search-engines-work/

[32] Ackison, M. (2021). SQL Tutorial: Guideline For Beginners: W3Schools Sql. Independently Published.

[33] Spinelli, J. (2018, December 19). MVC overview. Medium. https://medium.com/@joespinelli_6190/mvc-model-view-controller-ef878e2fd6f5

[34] Download XAMPP. (n.d.). Apachefriends.Org. Retrieved April 30, 2022, from https://www.apachefriends.org/download.html

[35] XAMPP tutorial: installation and first steps. (n.d.). IONOS Digitalguide. Retrieved April 30, 2022, from https://www.ionos.com/digitalguide/server/tools/xampp-tutorial-create-your-own-local-test-server/

[36] Composer. (n.d.). Getcomposer.Org. Retrieved April 30, 2022, from https://getcomposer.org/download/

[37] MySQL :: Download MySQL installer. (n.d.). Mysql.Com. Retrieved April 30, 2022, from https://dev.mysql.com/downloads/windows/installer/8.0.html

[38] Tìm hiểu về mô hình MVC trong Laravel. (2020, July 28). Thành Đỗ | Đào tạo Tin học trực tuyến. https://dothanhspyb.com/tim-hieu-ve-mo-hinh-mvc-trong-laravel/