**Name-Mr. Jain Samkitkumar Hasmukhlal**

**Roll No.-20 BEIT**

**Title- Write a case study on Web search engine.**

**WEB SEARCH ENGINE**

**INTRODUCTION:**

A **web search engine** is a software system that is designed to search for information on the [World Wide Web](https://en.wikipedia.org/wiki/World_Wide_Web). The search results are generally presented in a line of results often referred to as [search engine results pages](https://en.wikipedia.org/wiki/Search_engine_results_page) (SERPs). The information may be a mix of [web pages](https://en.wikipedia.org/wiki/Web_page), images, and other types of files. Some search engines also [mine data](https://en.wikipedia.org/wiki/Data_mining) available in [databases](https://en.wikipedia.org/wiki/Database) or [open directories](https://en.wikipedia.org/wiki/Web_directory). Unlike [web directories](https://en.wikipedia.org/wiki/Web_directories), which are maintained only by human editors, search engines also maintain [real-time](https://en.wikipedia.org/wiki/Real-time_computing) information by running an [algorithm](https://en.wikipedia.org/wiki/Algorithm) on a [web crawler](https://en.wikipedia.org/wiki/Web_crawler).

A **Web crawler** is an [Internet bot](https://en.wikipedia.org/wiki/Internet_bot) which systematically browses the [World Wide Web](https://en.wikipedia.org/wiki/World_Wide_Web), typically for the purpose of [Web indexing](https://en.wikipedia.org/wiki/Web_indexing) (*web spidering*).

[Web search engines](https://en.wikipedia.org/wiki/Web_search_engine) and some other sites use Web crawling or spidering software to update their [web content](https://en.wikipedia.org/wiki/Web_content) or indices of others sites' web content. Web crawlers can copy all the pages they visit for later processing by a search engine which [indexes](https://en.wikipedia.org/wiki/Index_%28search_engine%29) the downloaded pages so the [users](https://en.wikipedia.org/wiki/User_%28computing%29) can search much more efficiently.

Crawlers consume resources on the systems they visit and often visit sites without tacit approval. Issues of schedule, load, and "politeness" come into play when large collections of pages are accessed. Mechanisms exist for public sites not wishing to be crawled to make this known to the crawling agent. For instance, including a [robots.txt](https://en.wikipedia.org/wiki/Robots.txt) file can request [bots](https://en.wikipedia.org/wiki/Software_agent) to index only parts of a [website](https://en.wikipedia.org/wiki/Website), or nothing at all.

As the number of pages on the internet is extremely large, even the largest crawlers fall short of making a complete index. For that reason search engines were bad at giving relevant search results in the early years of the World Wide Web, before the year 2000. This is improved greatly by modern search engines, nowadays very good results are given instantly.

**HISTORY:**

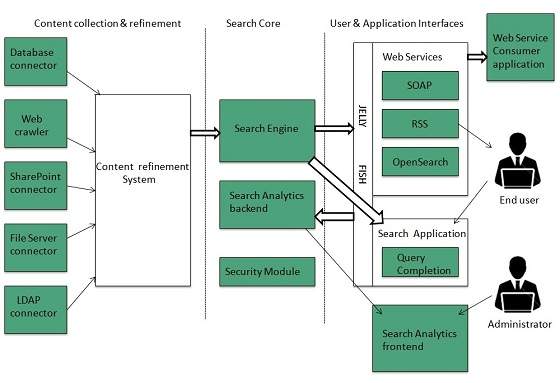
Internet search engines themselves predate the debut of the Web in December 1990. The [Who is](https://en.wikipedia.org/wiki/Whois) user search dates back to 1982 [[1]](https://en.wikipedia.org/wiki/Web_search_engine#cite_note-1) and the [Knowbot Information Service](https://en.wikipedia.org/wiki/Knowbot_Information_Service) multi-network user search was first implemented in 1989.[[2]](https://en.wikipedia.org/wiki/Web_search_engine#cite_note-2) The first well documented search engine that searched content files, namely [FTP](https://en.wikipedia.org/wiki/FTP) files was Archie, which debuted on 10 September 1990. Prior to September 1993 the World Wide Web was entirely indexed by hand. There was a list of [webservers](https://en.wikipedia.org/wiki/Webserver) edited by [Tim Berners-Lee](https://en.wikipedia.org/wiki/Tim_Berners-Lee) and hosted on the [CERN](https://en.wikipedia.org/wiki/CERN) webserver. One historical snapshot of the list in 1992 remains,[[3]](https://en.wikipedia.org/wiki/Web_search_engine#cite_note-3) but as more and more web servers went online the central list could no longer keep up. On the [NCSA](https://en.wikipedia.org/wiki/National_Center_for_Supercomputing_Applications) site, new servers were announced under the title "What's New!"[[4]](https://en.wikipedia.org/wiki/Web_search_engine#cite_note-4)

The first tool used for searching content (as opposed to users) on the [Internet](https://en.wikipedia.org/wiki/Internet) was [Archie](https://en.wikipedia.org/wiki/Archie_search_engine).[[5]](https://en.wikipedia.org/wiki/Web_search_engine#cite_note-LeidenUnivSE-5) The name stands for "archive" without the "v". It was created by [Alan Emtage](https://en.wikipedia.org/wiki/Alan_Emtage), Bill Heelan and J. Peter Deutsch, computer science students at [McGill University](https://en.wikipedia.org/wiki/McGill_University) in [Montreal](https://en.wikipedia.org/wiki/Montreal). The program downloaded the directory listings of all the files located on public anonymous FTP ([File Transfer Protocol](https://en.wikipedia.org/wiki/File_Transfer_Protocol)) sites, creating a searchable database of file names; however, [Archie Search Engine](https://en.wikipedia.org/wiki/Archie_search_engine) did not index the contents of these sites since the amount of data was so limited it could be readily searched manually. In the summer of 1993, no search engine existed for the web, though numerous specialized catalogues were maintained by hand. [Oscar Nierstrasz](https://en.wikipedia.org/wiki/Oscar_Nierstrasz) at the [University of Geneva](https://en.wikipedia.org/wiki/University_of_Geneva) wrote a series of [Perl](https://en.wikipedia.org/wiki/Perl) scripts that periodically mirrored these pages and rewrote them into a standard format. This formed the basis for [W3Catalog](https://en.wikipedia.org/wiki/W3Catalog), the web's first primitive search engine, released on September 2, 1993.[[6]](https://en.wikipedia.org/wiki/Web_search_engine#cite_note-Announcement_html-6)

In June 1993, Matthew Gray, then at [MIT](https://en.wikipedia.org/wiki/Massachusetts_Institute_of_Technology), produced what was probably the first [web robot](https://en.wikipedia.org/wiki/Web_robot), the [Perl](https://en.wikipedia.org/wiki/Perl)-based [World Wide Web Wanderer](https://en.wikipedia.org/wiki/World_Wide_Web_Wanderer), and used it to generate an index called 'Wandex'. The purpose of the Wanderer was to measure the size of the World Wide Web, which it did until late 1995. The web's second search engine [Aliweb](https://en.wikipedia.org/wiki/Aliweb) appeared in November 1993. Aliweb did not use a [web robot](https://en.wikipedia.org/wiki/Web_robot), but instead depended on being notified by website administrators of the existence at each site of an index file in a particular format. Soon after, many search engines appeared and vied for popularity. These included [Magellan](https://en.wikipedia.org/wiki/Magellan_%28search_engine%29), [Excite](https://en.wikipedia.org/wiki/Excite), [Infoseek](https://en.wikipedia.org/wiki/Infoseek), [Inktomi](https://en.wikipedia.org/wiki/Inktomi_%28company%29), [Northern Light](https://en.wikipedia.org/wiki/Northern_Light_Group), and [AltaVista](https://en.wikipedia.org/wiki/AltaVista). [Yahoo!](https://en.wikipedia.org/wiki/Yahoo%21) was among the most popular ways for people to find web pages of interest, but its search function operated on its [web directory](https://en.wikipedia.org/wiki/Web_directory), rather than its full-text copies of web pages. Information seekers could also browse the directory instead of doing a keyword-based search. [Google](https://en.wikipedia.org/wiki/Google) adopted the idea of selling search terms in 1998, from a small search engine company named [goto.com](https://en.wikipedia.org/wiki/Yahoo%21_Search_Marketing). This move had a significant effect on the SE business, which went from struggling to one of the most profitable businesses in the internet

Around 2000, [Google's search engine](https://en.wikipedia.org/wiki/Google_Search) rose to prominence.[[11]](https://en.wikipedia.org/wiki/Web_search_engine#cite_note-11) The company achieved better results for many searches with an innovation called [PageRank](https://en.wikipedia.org/wiki/PageRank), as was explained in the paper *Anatomy of a Search Engine* written by [Sergey Brin](https://en.wikipedia.org/wiki/Sergey_Brin) and [Larry Page](https://en.wikipedia.org/wiki/Larry_Page), the later founders of Google.[[12]](https://en.wikipedia.org/wiki/Web_search_engine#cite_note-12) This [iterative algorithm](https://en.wikipedia.org/wiki/Iterative_algorithm) ranks web pages based on the number and PageRank of other web sites and pages that link there, on the premise that good or desirable pages are linked to more than others. Google also maintained a minimalist interface to its search engine. In contrast, many of its competitors embedded a search engine in a [web portal](https://en.wikipedia.org/wiki/Web_portal). In fact, Google search engine became so popular that spoof engines emerged such as [Mystery Seeker](https://en.wikipedia.org/wiki/Mystery_Seeker).

**Architecture**:



The search engine architecture comprises of the three basic layers listed below:

* Content collection and refinement.
* Search core
* User and application interfaces

**Search Engine Processing**

**Indexing Process**

Indexing process comprises of the following three tasks:

* Text acquisition
* Text transformation
* Index creation

**Text acquisition**

It identifies and stores documents for indexing.

**Text Transformation**

It transforms document into index terms or features.

**Index Creation**

It takes index terms created by text transformations and create data structures to suport fast searching.

### Query Process

Query process comprises of the following three tasks:

* User interaction
* Ranking
* Evaluation

#### User interaction

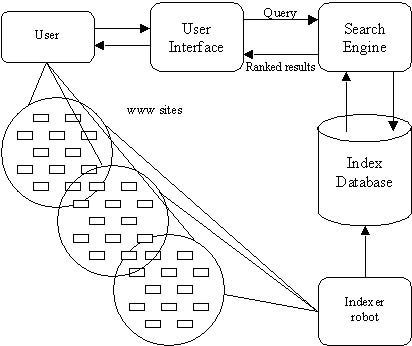
It supports creation and refinement of user query and displays the results.

#### Ranking

It uses query and indexes to create ranked list of documents.

#### Evaluation

It monitors and measures the effectiveness and efficiency. It is done offline.



**GUI and OUTPUT:**



Fig: GUI of search engine

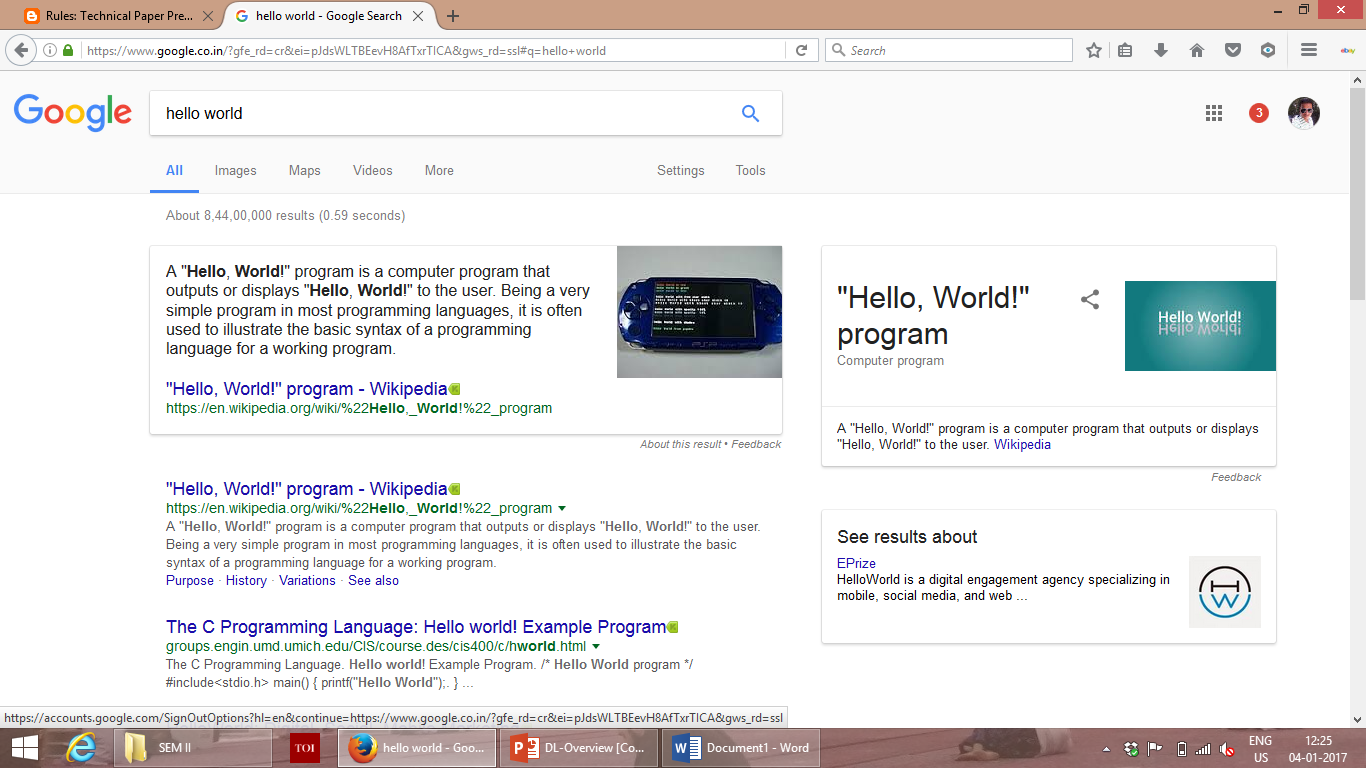


Fig: Output of search engine

**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. Without a search engine, you would have to look at sites one by one and pore over the contents of each carefully – a tedious prospect. A search site automatically compares your criteria to billions of Web pages and gives you results in a fraction of a second. You can perform dozens of searches in the course of a few minutes, altering the criteria as you narrow down results.

## Relevance:

## When a search engine scans a website, it scores the content for relevance to particular search words. For example, the site, "Joe's Pizza" scores higher for the search term "pizza restaurant" than the site "Acme Pizza Dough Distributors." Although search companies keep their ranking formulas secret, they take into consideration factors such as repetition of related words and links from other sites. A search engine sorts its results page by relevance to your criteria, with the score in descending order. You see the highest-scoring results at the top; as you move down the list, websites become less relevant.

## Free Access:

## Some search engines, such as LexisNexis, specialize in legal or other specialized, scholarly information; these sites charge a fee to use their services. Google, Bing and Yahoo pay for their operations through advertising; searches are free to the user, without restrictions for the information you seek, the time spent on the site or the number of searches you perform. Although this benefits all users, it is a particular advantage for students, job seekers and others of limited means.

## Comprehensive:

## Search engines scan the entire Web and keep comprehensive data on every page they catalog. Because search companies hold so much data, they help you find obscure sites about which you would not otherwise know. Search results are more likely to give you too much information rather than too little.

## Advanced Search:

## In addition to keywords, search engines let you use advanced search options to refine your results. These options help make your searches more flexible and sophisticated. For example, to exclude results containing a certain word, type a minus sign before the word. To look for an exact phrase, surround it with quotation marks. When you want to search only a specific site or group of sites, type "site:" without quotes followed by the site's Web address.

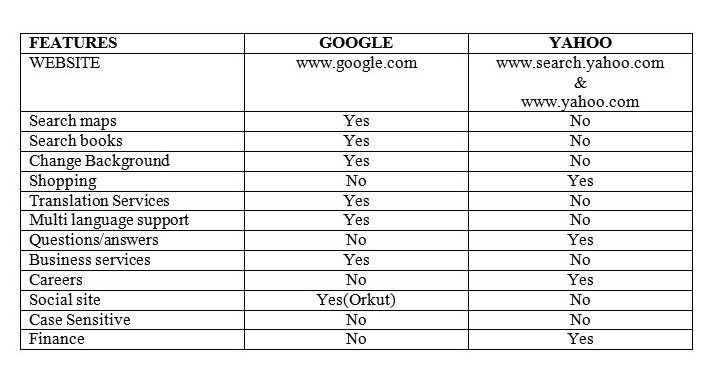
## DISADVANTAGES:

## Regardless of the growing sophistication, many well thought-out search phrases produce list after list of irrelevant web pages. The typical search still requires sifting through dirt to find the gems.

## Using search engines does involve a learning curve. Many beginning Internet users, because of these disadvantages, become discouraged and frustrated.

1. They rank websites in part according to concepts such as intrinsic **authority**, which in many cases are flawed, and which allow rankings to be manipulated.
2. They **filter** results according to the information that a particular user has given them, which rarely provides an accurate reflection of a user’s real interests.
3. Although it is less of a problem than it used to be, search engine rankings have always been manipulated by **keyword–stuffing**: the inclusion of unnaturally large numbers of search terms into web pages.
4. Rankings are almost certainly affected in some way by the search engine companies’ commercial interests: their dependance on **paid advertising**, and their promotion of their **own products**.

**COMPARISON:**



**CONCLUSION:**

Thus we made a detailed study of working, architecture, advantages, disadvantages of various search engines available in market. We also made a comparative study of two widely known search engines i.e. ‘Yahoo’ and ‘Google’ based on various parameters.