*Romanian Journal of Petroleum & Gas Technology*
*VOL. IV (LXXV) • No. 2/2023*

# AN ANALYSIS OF INFORMATION SEARCH AND RETRIEVAL TECHNIQUES

**Otilia Cangea** [1]

[1] Petroleum-Gas University of Ploiesti, Romania
email: ocangea@upg-ploiesti.ro

**DOI: 10.51865/JPGT.2023.02.14**

## ABSTRACT

The informational growth dramatically changed the standard form of publication, access, and use of scientific information. Due to Internet, scientific publications migrated from the traditional to a digital configuration. Common users, as well as highly trained professionals from various fields, aim to retrieve information in the shortest possible time, but in the context of an unprecedented growth of data volume, big difficulties occur in finding and accessing the available information, resulting in stress and important delays in decision making. Therefore, training courses in information culture are organized in university libraries, particularly for a superior efficiency of scientific data bases usage. The paper presents an introduction to the architecture of a retrieval information system and a study of the most popular search engines and of the indexing process. Hereinafter, a Page Rank algorithm implementation is performed and a comparative study relative to different search engines is conducted, emphasizing the specific advantages and disadvantages.

**Keywords**: information retrieval, search engines, Page Rank algorithm

## INTRODUCTION

Stored in data bases with different structures and functionalities, the information is accessible even without properly knowing the respective searched bases [4]. The process is difficult by reason of the big number of searching tools, the various information and methods content, and the lack of standards; all of these in the circumstances of a significant increase, in the last decade, of daily Internet users, as seen in figure 1.

Since there are a multitude of information accessible on web, these must be organized. Furthermore, instruments or programs to help locating the information are needed. There are many sources of information and search instruments on Internet; the most important are:

- Web Directory for Internet general resources collection, depending on the subject;
- Virtual libraries – directory or catalog of subjects with selected web resources;
- Specialized data bases with hyperlinks for a specific domain or indexes accessible on web;
- FTP archives – file collections in various forms available on Internet;
- Search engines dependent on key words;
- Metasearch instruments that allow access to data bases.

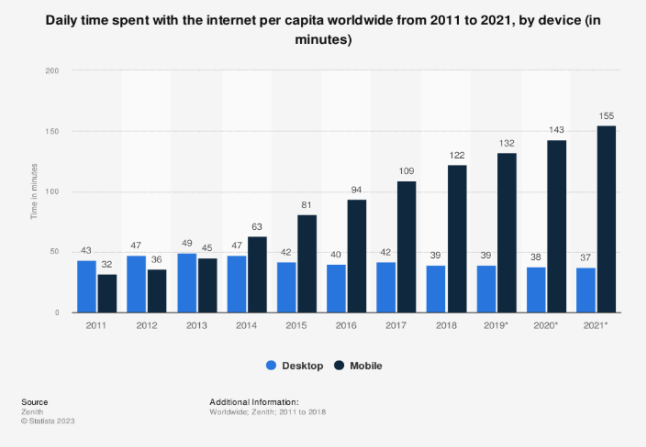


***Figure 1.*** *Daily Internet usage per capita worldwide 2011-2021 [5]*

An example that displays the result of navigation in *Engineering and Technology* directory (https://www.einet.net/) and *Computer Technology/Algorithms* subdirectories is presented in figure 2.


***Figure 2.*** *Algorithms range in eiNET directory*

An example of a references data base is UCLA Library (University of California, Los Angeles), free of charge available at https://catalog.library.ucla.edu/vwebv/searchBasic. The first page presents a key word search, the attached restrictions and the number of recordings (figure 3).

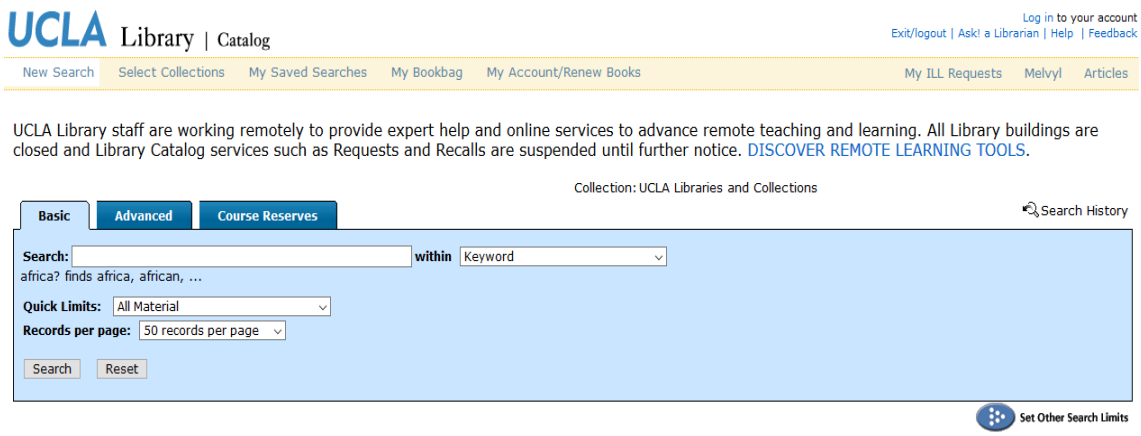


***Figure 3.*** *First page of UCLA Library*

In the context of information retrieval and web searching, one can define the retrieval model of a web document by three elements: the representation pattern for a document, interrogation, and the relevance/rating of a document function.

From the web classification viewpoint, there are five principal representation models [7] and all are based upon the same principle (Salton, 1971): each document and interrogation are considered as a *bag of words*. These models, namely:

- Boolean model (Lancaster & Fayen, 1973);
- Vector Space Model, (VSM) (Salton, 1971);
- Linguistic model (Ponte & Croft, 1998);
- Probabilistic model (Jones, et al., 2000);
- Support Vector Machines model, (SVM) (Vapnik, 1995).

define the so-called *layout representation* concept.

Considering these, an information retrieval system is defined by the representation form of a document, respectively of an interrogation, and by the function that valuates the relevance document-interrogation. The architecture of a retrieval system is presented in figure 4.

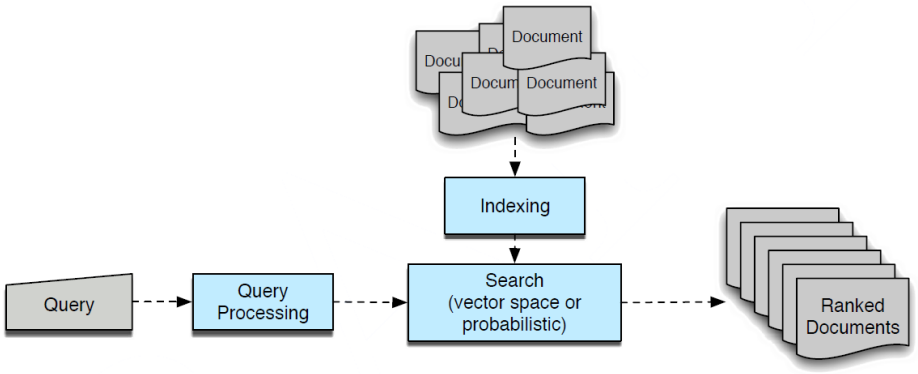


***Figure 4.*** *Architecture of a retrieval system [4]*

## SEARCH ENGINES

A web search engine is a software system designed for searching and identifying items in a database corresponding to specific keywords used for finding particular sites on the web. As distinct from web directories, managed exclusively by human editors, search engines maintain information in real time by running a web crawling algorithm [1].

Historically speaking, search engines on the Internet appeared after 1980; the *WHOIS* search/answer protocol was implemented in 1982, and *Knowbot Information Service***,** considered the first search engine, in 1989, although it searched users, not information content. The first real approved search engine was *Archie* (derived from the word *archive*), launched in September 1990, came into use from the necessity of knowing all World Wide Web sites [3].

Year 1994 marked the appearance of the first modern search engines generation *WebCrawler*, *AltaVista*, *Infoseek*, *Lycos*, *Excite*, *Ask Jeeves*, and the famous *Yahoo!*. WebCrawler, still in use, was the first to entirely scan web pages and allowed users to make keywords search [3].

In September 1996 *Google* was launched; the initial design of the BackRub experiment by Larry Page and Sergey Brin constantly grew, so that in 2002 it became the most popular and remained on the first position according to a study conducted in August 2023: a share of 91.85% of the entire search engine market comes to *Google* (figure 5) [11].

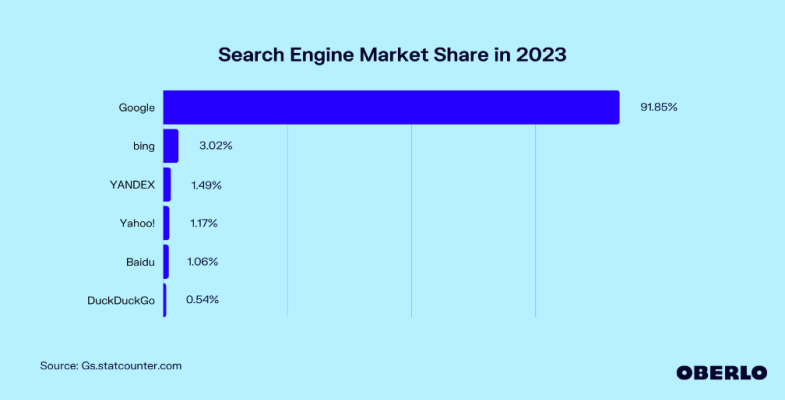


***Figure 5.*** *Search engine market in 2023 [11]*

Another important search engine is **MSN** (Microsoft Network), that uses as back end the previously obtained results of the pre-existent engines Inktomi and Look Smart; beginning with 2004 Microsoft uses its own back end for indexing (figure 6) [4].

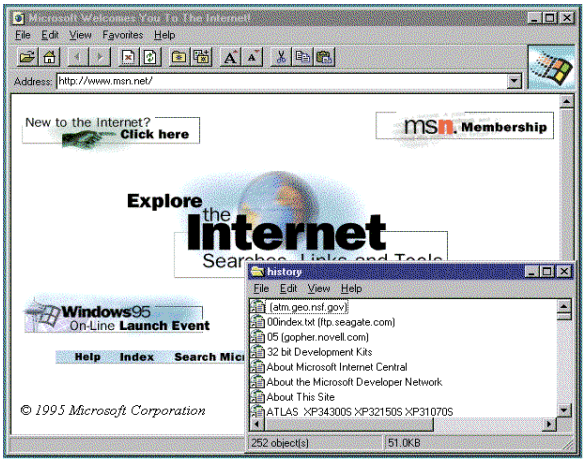


***Figure 6.*** *MSN search engine [4]*

*WEB CRAWLER*

The essential component of every search engine is the web crawler, with an architecture presented in figure 7. A web crawler collects the new documents from the *www* virtual space; it goes over the web pages, analyzes their content using a specific algorithm, builds data structures for a subsequent processing, and follows the links to other pages. This process is managed by the *Controller* or *Scheduler*, responsible for supervising the threads. (vulnerabilities). Each page, identified by URL address, is downloaded from web and transferred to *Lexer* for a lexical analysis. *Content Processor* then builds the required

indexing and other specific tasks. Finally, a search in content referring to URL addresses transmitted to *Controller* is performed.

There are many applications for web crawlers; one of the most important is monitoring web pages so that a specific user or an entire community can be notified when new interesting information occur [8].

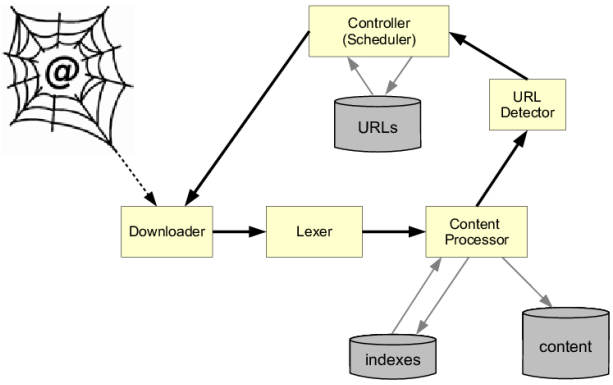


***Figure 7.*** *Web Crawler architecture [6]*

The most common crawler web-type algorithm starts with a set of URL addresses and uses the links found in these web pages to reach other pages. The process is recurrent until a sufficient number of pages are visited or another goal is reached [4]. Google founders, Serghei Brin and Lawrence Page, in *The anatomy of a large-scale hypertextual Web search engine* [2], identified the Web crawler as the most sophisticated, yet fragile, component of a search engine.

*INDEXING*

Additional to web crawler, another essential feature of a search engine is the indexing process; after detecting a page or a document, all gathered data are kept on the servers to be further indexed. The index of Google search contains hundreds of billion web pages and has a dimension of over 100.000.000 GB. Figure 8 presents the main stages of the indexing process; during the analysis, links are extracted to build a knowledge graphic that can be later analyzed to generate scores.

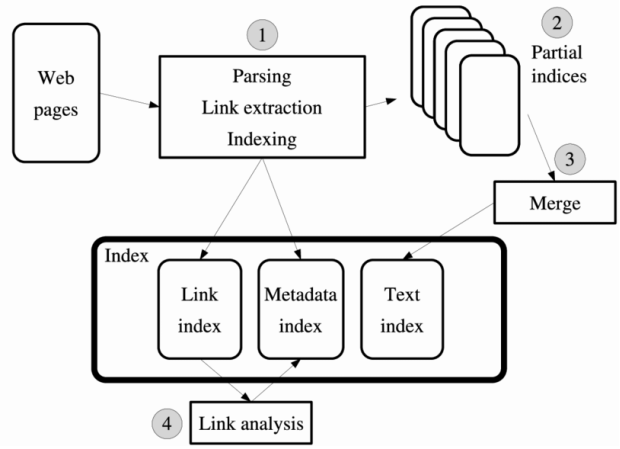


***Figure 8.*** *Indexing mechanism performance schema [3]*

A brief description of the indexing process, presented in figure 8, is as follows:

(1) pages are processed and analyzed, links are extracted and indexed;

(2) partial indexes are written on the disk when the main memory is exhausted;

(3) indexes are merged into a complete text index;

(4) an analysis of the off-line links is performed to calculate the static scores.

## PAGE RANK ALGORITHM IMPLEMENTATION

Google search engine is based on many representations that contain hyperlinks; these are basic representations for rapid calculation of the Page Rank coefficient for each web page. *PageRank* (PR) is an algorithm for the analysis of hyperlinks in Internet that assigns a weight to each element from an ensemble of documents interconnected by hyperlinks to measure the relative importance within the ensemble. Thus, if page A contains a link to page B, one can presume by default that A states about B that this is important, so B has to be better rated in the classification. The more qualitative links to a site, the greater the PR value and the better the classification. PR assigned by page B to page A decreases proportional to the number of links on page B. PR of page A is [2]:

$$PR(A) = (1 - d) + d\left(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right) \quad (1)$$

where:

- $d$ is a parameter (damping factor) set between 0 and 1 (usually 0.75);
- $T_1 \ldots T_n$ are pages (hyperlinks) to the page for which one calculates the PR;
- C ( $T_i$ ) is defined as the number of links of $T_i$ page;
- PR($T_i$->A) is the PR of $T_i$ pages that contain a link to A.

In this context, *www* is represented as a network, where web sites are nodes and the connections between them are edges. For example, a simplified configuration has only four web sites; let these be *A*, *B*, *C*, and *D*. After installing Networkx, one can run the scripts that generate the four nodes. Further, one must add the links between the nodes to generate the links matrix that describes, its interior and exterior links. If a node has *k* output edges, it will transmit an *1/k* fraction from its weight to each of the linked nodes, so that the following matrix is generated.

```
>>> A=np.matrix([(0,0,0,1),(1,0,0,0),(0,1,0,0),(0,0,1,0)])
>>> A
matrix([[0, 0, 0, 1],
        [1, 0, 0, 0],
        [0, 1, 0, 0],
        [0, 0, 1, 0]])
```

By an iterative process, one can update the rank vector *v* according to A matrix; the updated probability of the sites after an iteration will be vector *v '*, *v '= Av* and after *k* iterations is:

$v_1 = Av_0$

$v_2 = A^2v_0 = A(Av_0) = Av_1$

...

$v_k = A^kv_0 = Av_{k-1}$

Using Page Rank algorithm:

```
>>> pr=nx.pagerank(DG_test,alpha=1)
>>> pr
{1: 0.38709615908859496, 2: 0.12903204605249047, 3: 0.29032302109901886, 4: 0.193548773759895}
>>>
```

the obtained results (k = 1000) for rank vector updating are:

```
>>> np.array((B**1000)*A.T)
array([[0.38709677, 0.38709677, 0.38709677, 0.38709677],
       [0.12903226, 0.12903226, 0.12903226, 0.12903226],
       [0.29032258, 0.29032258, 0.29032258, 0.29032258],
       [0.19354839, 0.19354839, 0.19354839, 0.19354839]])
```

Figure 9 displays the values calculated using PR for a simple network. C page has a bigger PR than page E, although less links to it – the link from page B has a greater value. For example, a person navigating on web that randomly chooses a link from each page (with 15% probability to jump to a random page) will reach page E with 8,1% probability; the same rule applies for the other pages.

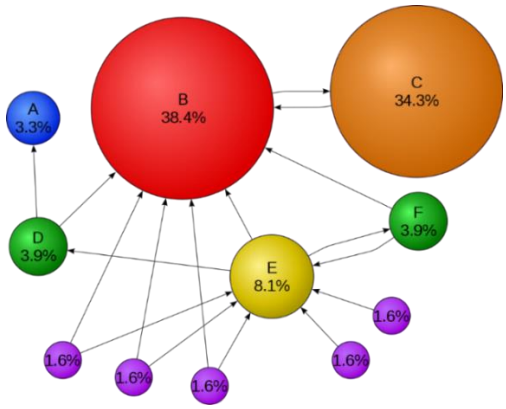


***Figure 9.*** *Values calculated using PageRank for a simple network*

For a more complex network, with 10 nodes

```
>>> G=nx.fast_gnp_random_graph(10,0.5,directed=True)
>>> nx.draw(G,with_labels=True)
>>> plt.show()
>>>
```

the result is presented in figure 10.

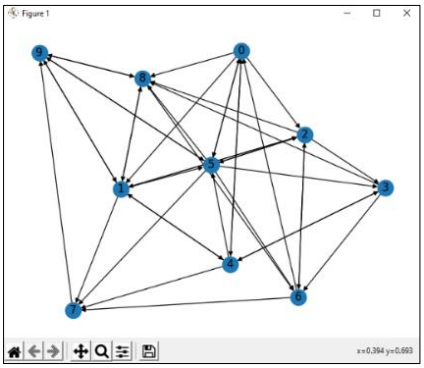

***Figure 10.*** *Example of PR implementation for a 10 nodes network*

## COMPARATIVE ANALYSIS OF SEARCH ENGINES

*GOOGLE*

Using Google search engine is fairly intuitive. One must introduce a key word or a group of words related to the desired result and the answer is offered as a list of hyperlinks to websites that meet the respectively words. For example, if we want to find information about C++ language programming, one can search the results using the words "learn C++ programming"; the results are displayed in figure 11. Google provides the users a multitude of facilities for advanced search. Other than the predefined options of format changing, such as *Video clips*, *Images*, *News*, there are key words used to add filters, such as the *filetype* key word [7]. Moreover, if only pdf format results are required, one can use the combination: *learn c++ programming filetype:pdf*, with *filetype* filter (figure 12)

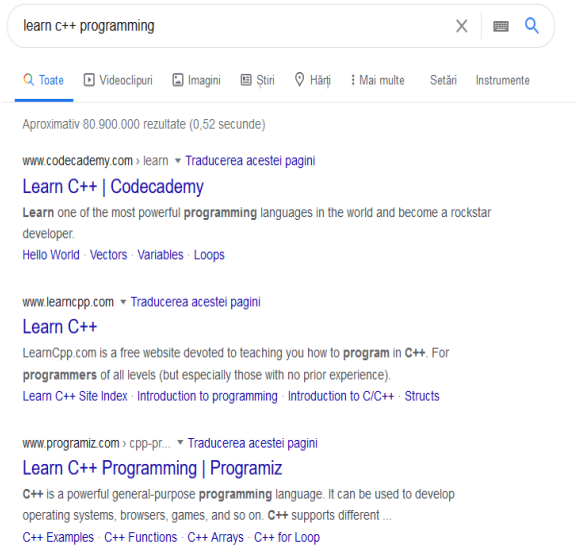


***Figure 11.*** *Simple search using Google engine*

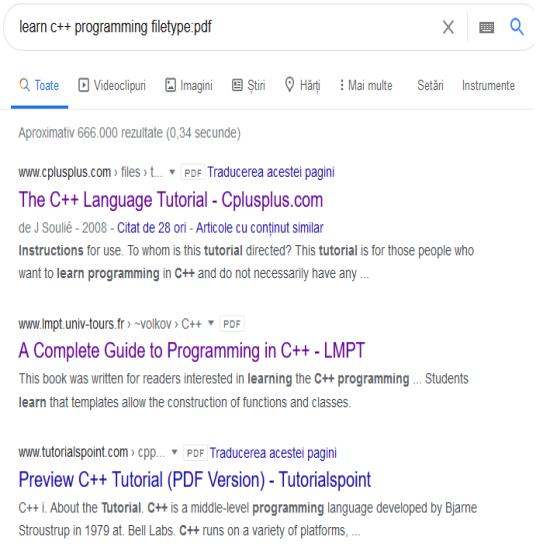


***Figure 12.*** *Search using filetype filter*

The search engine can also be used for performing various calculations or solving equations, as well as for graphical representations (figure 13, figure 14.)

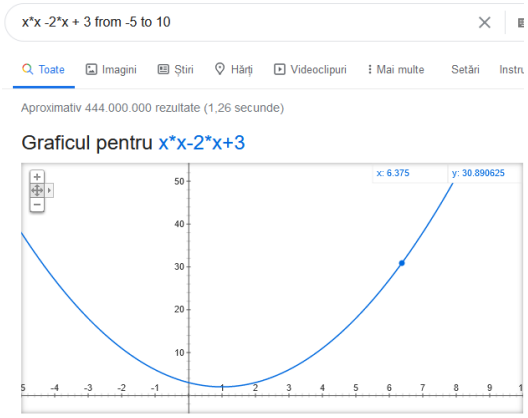


***Figure 13.*** *Graphical representation*

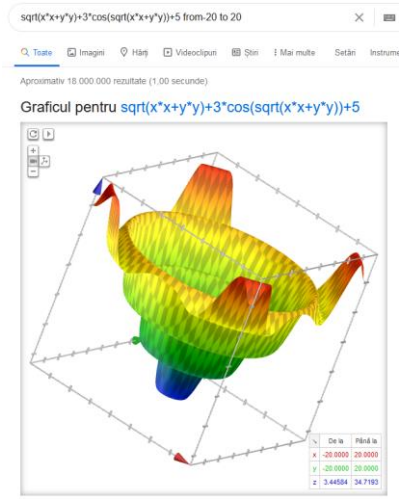


***Figure 14****. 3D graph for a complex relation*

The result of a *Featured Snippet* search is the answer that Google considers to be the best (figure 15).

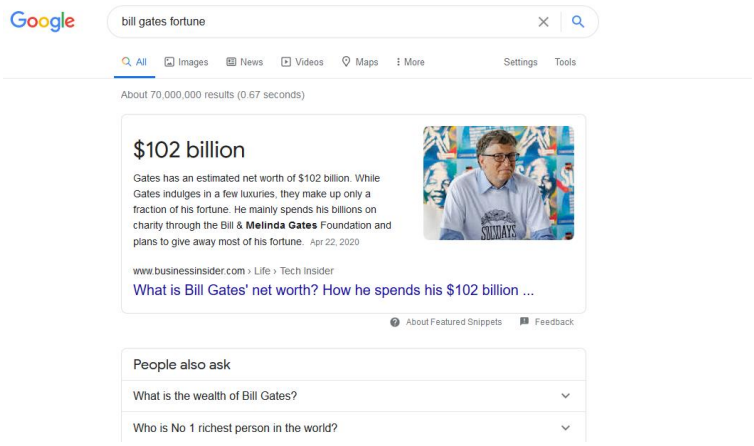


***Figure 15**. Example of a Featured Snippet search for 'bill gates fortune'*

*Advantages*

- the best search index and the most advanced algorithms;
- dominant power in market; according to *statcounter.com,* Google holds 92.71% from global searches, next being Microsoft Bing (appx. 3% )[8].

*Disadvantages*

- invasive advertising and personal data collection, in contradiction with GDPR legislation;
- frequent update of interfaces and algorithms; Google often tests new functionalities, thus generating difficulties for users and developers.

*DUCKDUCKGO*

Founded in 2008 [17], **DuckDuckGo** (DDG) company respects confidentiality – it does not memorize and use any personal data, save search historic or follow IP addresses. The user interface e has numerous similarities with Google:

- the search key words are written at the right side of the logo (figure 16);

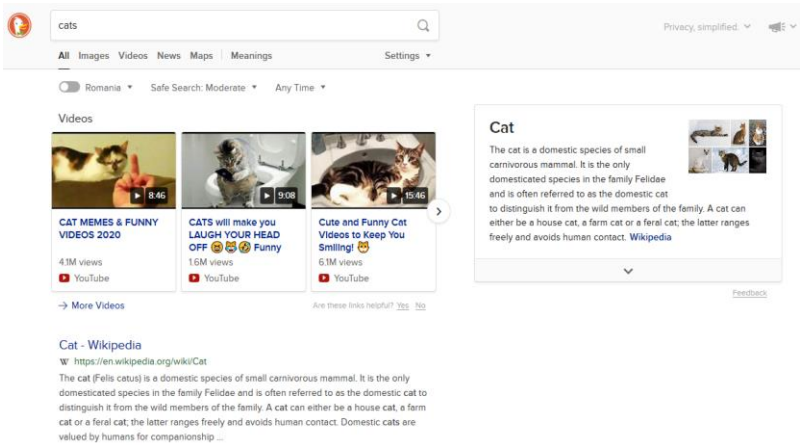


***Figure 16.** Search using "cats" keyword in DDG*

- the search results are embedded in SERP (Search Engine Results Pages): DDG displays 10 results on the first page, and Google only 8; for more pages, DDG uses 'More results' button (figure 18), and Google displays the number of pages with the possibility of click on each page (figure 17).

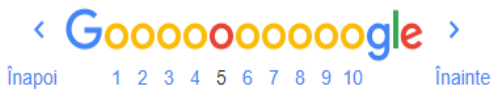





***Figure 17.*** *Accessing more pages in Google* ***Figure 18.*** *Accessing more pages in DDG*

- the results, similar to *Featured Snippet* in Google, are displayed on the right side of the page;
- uses published reviews together with addresses, telephone numbers or work schedules, and the directions for maps are given by Bing Maps set by default (figure 19)

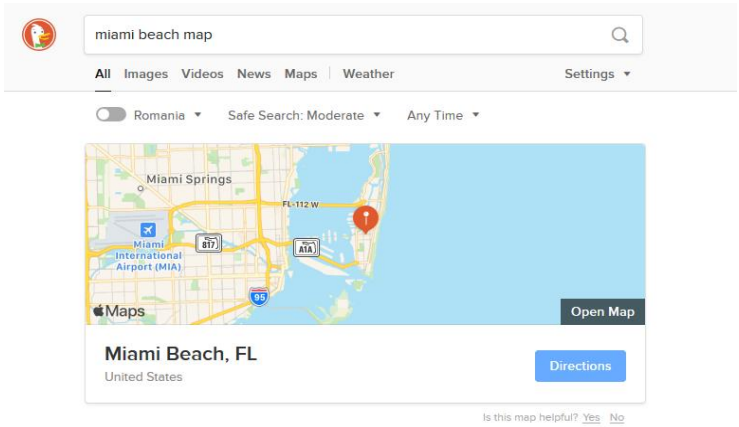


***Figure 19****. Map for Miami beaches (DDG using Bing Maps)*

*Advantages*

- an increasing popularity – more users, more profit;
- ease in use; friendly interface, an easy method to information search;
- privacy – it does not follow information referring to user personal data or their search, claiming that offer the most secure private search.

*Disadvantages*

- a market share of 0.49 % (October 2022), much smaller than Google and even smaller than Microsoft Bing, Yahoo or Baidu [8];
- for specific regional searches, the content is provided in English.

*MICROSOFT BING*

Microsoft Bing, launched in 2009, originates from MSN Search, Windows Live Search, and Live Search [9]. Comparing to Google, Bing offers the same basic functionalities: text, video, images, and maps search, with a similar page display (figure 20). As for the speed and content, Google offered 108 000 000 results in 0.95 seconds and Bing, only 6 430 000 results.

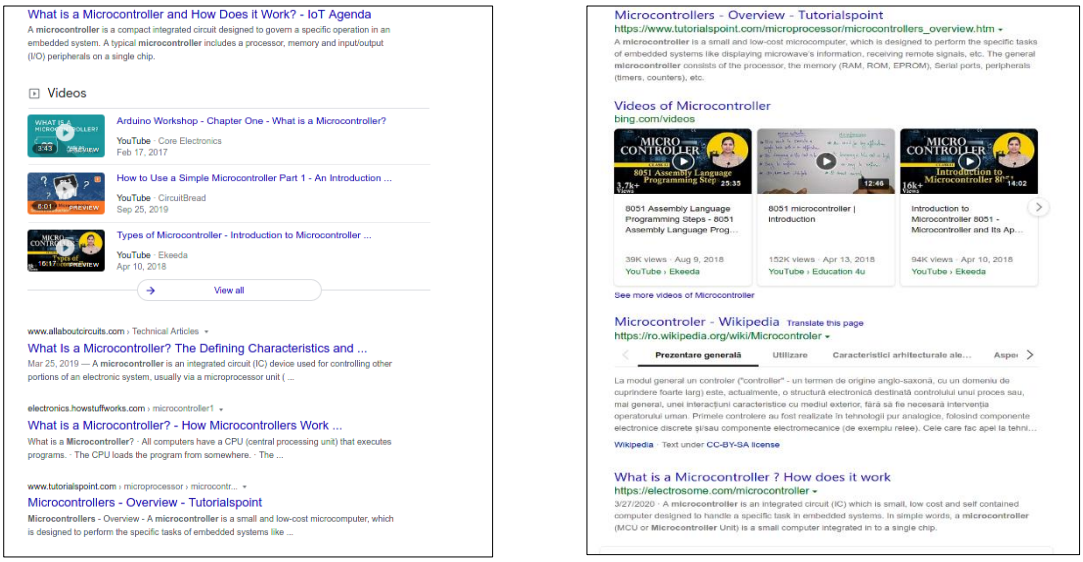


***Figure 20.*** *Search results for „microcontrollers" using Google and Microsoft Bing*

*Advantages*

- image quality better than with Google;
- a Microsoft Rewards program, obtaining reward points for discounts, access to other products, donations to charity organizations;
- design of first page – image from nature, compared to the simple Google logo;
- more auto search suggestions than Google.

*Disadvantages*

- less results with reduced relevance;
- a very reduced usage segment on the search engines market;
- limited geographical presence, the majority of the users being from USA.

## CONCLUSIONS

The most popular search engines are Google, Yahoo and Bing. In the majority of cases, an Internet search will lead to a very long list of results; hence, the search will display millions of web pages.

Today, website navigation is very familiar; Internet offers mainly the following services:

- information search using dedicated search engines
- information retrieval
- chat, newsgroup, forum
- e-mail, e-commerce
- banking transactions
- on-line databases,

with information retrieval depending on:

- data source, data bases
- specific methods for databases management
- search engines.

The comparative study specifies the advantages and disadvantages of the use of online search engines; some of the most important are: time saving, relevance, free access, complete information content, and advanced search.

Future research directions may consider designing a more performant web browser that can search and retrieve information on Internet in a minimum period of time.

**REFERENCES**

1. Ntoulas A., Cho J., Olston C., What's New on the Web? The Evolution of the Web from a Search Engine Perspective, Proceedings of International Conference on World Wide Web, 2004
2. Brin, S., Page, L., The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems, Volume 30, Issues 1–7, pp. 107-117, 1998
3. Castillo, C., Effective Web Crawling, Dept. of Computer Science, University of Chile, 2004
4. Jurafsky, D., James H. M., Question Answering and Summarization, Chapter 23 Speech and Languages Processing, Second Edition, Prentice-Hall, Inc., 2009
5. Petrosyan, A., Daily internet usage per capita worldwide 2011-2021, Aug 25, 2023, available at: https://www.statista.com/statistics/319732/daforily-time-spent-online-device/
6. Turek, W., Nawarecki, E., Dobrowolski, G., Krupa, T., Majewski, P., Web Pages Content Analysis using Browser-based Volunteer Computing, Computer Science 14(2), DOI: 10.7494/csci.2013.14.2.215, 2013
7. Zhai, C., Lafferty, J., A study of smoothing methods for language models applied to ad hoc information retrieval. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pg. 334-342, New Orleans, United States, 2001
8. *** What is a Web Crawler? How Spiders Work, Clouflare Learning center, available at: https://www.cloudflare.com/learning/bots/what-is-a-web-crawler/
9. *** The Evolution of Search. Understanding User Intent, SEOPressor, 2023, available at: https://seopressor.com/blog/evolution-of-search/
10. *** Search Engine Market Share Worldwide Sept. 2022-Sept. 2023, Global Stats, available at: https://gs.statcounter.com/search-engine-market-share
11. *** Search engine market share 2023, OBERLO Statistics, available at: https://www.oberlo.com/statistics/search-engine-market-share
12. http://www.bcu-iasi.ro/docs/biblos/biblos8/regasireainf.pdf
13. https://en.wikipedia.org/wiki/Timeline_of_web_search_engines
14. https://en.wikipedia.org/wiki/Google_Chrome

Received: November 2023; Accepted: November 2023; Published: November 2023