NOVATEUR PUBLICATIONS
JournalNX- A Multidisciplinary Peer Reviewed Journal
ISSN No: 2581 - 4230
VOLUME 6, ISSUE 6, June -2020

ANALYSIS OF ADVANCED SEARCH ENGINE MECHANISMS

MENGATOVA XURSHIDA TOSHMUXAMATOVNA
Teacher, Termez branch of Tashkent State Technical University named after Islam Karimov,
xurshidamengatova7288@gmail.com

ESONTURDIYEV MAMATKOBIL NURMAMATOVICH
Teacher, Chirchiq State Pedagogical Institute of Tashkent region,
esonturdiyev80@mail.ru

ABSTRACT:
This article presents an analysis of the mechanism of work of advanced search engines, which are currently developing and have a wide range of capabilities. Have been studied technologies for extracting text data from text files (txt), HTML pages (htm, html), Adobe Acrobat files (pdf), MS Word and MS Excel files (doc, docx, xls, xlsx), multimedia files (mp3, avi, mpeg, etc.) and executable files (exe) in search engines. The indexing processes were analyzed, as well as the stages of data search and sorting of results. The results of the analysis can be used to develop general functional requirements for a database of electronic documents when creating an advanced search system.
Keywords: Select files, text files (txt), HTML pages (htm, html), Adobe Acrobat files (pdf), MS Word and MS Excel files (doc, docx, xls, xlsx), multimedia files (mp3, avi, mpeg, etc.), executable files (exe), indexing, data search and sorting results.

INTRODUCTION:

Currently, as electronic processing of documents is growing rapidly, organizations are faced with a number of problems when working with large volumes of electronic documents stored on servers on the local network. As the volume and size of electronic documents increases, one of the most important tasks is the problem of their sorting, reliable storage, as well as ensuring their effective use.

The growing volume of electronic documents causes problems with finding the necessary information quickly and accurately. As a solution to this problem, many large companies working in the field of barbecue, presented their products. Work is also underway to further improve the flour.

The following is an analysis of the performance mechanism of advanced search engines, which is currently being developed and has a wide range of capabilities.

MAIN PART:

All modern software for search engines have their own algorithms for searching and sorting data, the principle of which is based on a common mechanism. This mechanism includes the following steps: file selection, indexing, data retrieval, and sorting of results[6]. A general functional diagram representing these steps is shown in Figure 1 below.

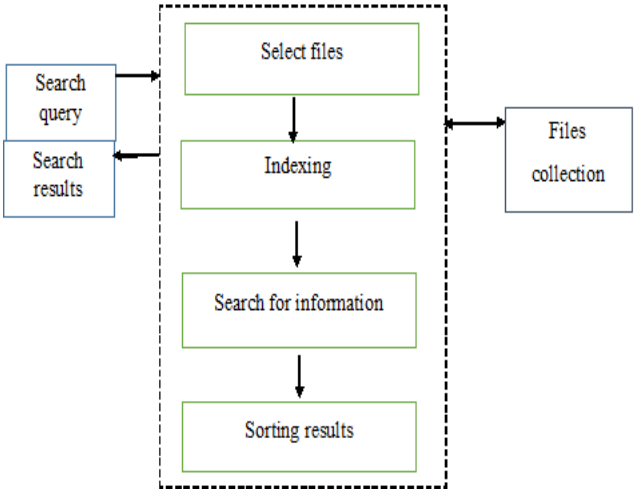

Figure 1. General functional diagram of advanced search engines

Below is an analysis of each stage of the advanced search engine engine.

SELECT FILES:

When search engines are launched, they start accessing the files located on the disk or the folders attached to them and select all the files that have been newly entered into it. These files can be text, image, video, and other types of files. The search engine constantly analyzes the information contained in files to check old files, change them if necessary and add new ones [6].

Search engine performance is based on the use of the information contained in the file, which means the ability to extract text words from each file. In most cases, the solution to this problem is simple (when the file is a text file), but there are files of such a format that it is difficult to separate text words from their contents. The following is an analysis of the files selected in the search engine and extracting text words from them.

Text Files (TXT).

Although this is the most convenient file for extracting text words, today it is rarely used.

HTML pages (htm, html):

Since the contents of these files are in a structured state, extracting text words from it is not a problem.

Adobe Acrobat Files (pdf):

Extracting text words from the contents of these files is much more difficult. There are a number of programs that perform a complete analysis of the text contained in these files. But converting these files to a text representation takes a lot of time.

MS Word and MS Excel files (doc, docx, xls, xlsx):

Although Microsoft has not officially disclosed the classification of these formats, there are unofficial classifications and a number of solutions on the Internet today for extracting text from these files [2].

Multimedia files (mp3, avi, mpeg, etc.):

They can be used to distinguish between album names, videos and music, if this information is included in them.

Executable files (exe):

Although it is impossible to save the text in these files, this format allows you to identify the program and its authors.

For files that do not have the ability to extract text directly, you can extract information such as its name, the user who created it [2].

INDEXING:

Indexing is used in search engines to quickly, accurately search for information and speed up the process of storing, sorting and selecting data. In the process of indexing, the files and the words that appear in them are assigned serial numbers - indexes. This information is stored in search engine indexes.
In turn, the indexing process is carried out in two different ways.:

- proper indexing;
- reverse indexing.

During proper indexing, a table is created that contains the words in each file and their sequence numbers. The reverse indexing table is formed on the basis of the number of files and files in which the word is present, corresponding to each word [1].

The indexing methods listed above can be seen in the following three sample files.
The content of file 1: “In many enterprises, electronic data is stored and processed in computer systems.”
The content of file 2: “Search engines work with data stored in the database and entered parameters.”

The content of file 3: "The physical model of the database is implemented using database management systems."

Table 1. Proper Indexing Table

File 1	in: 1,10; many: 2; enterprises: 3; electronic: 4; data: 5; is: 6; stored: 7; and: 8; processed: 9; computer: 11; systems: 12;
File 2	search: 1; engines: 2; work: 3; with: 4; data: 5; stored: 6; in: 7; the: 8; database: 9; and: 10; entered: 11; parameters: 12
File 3	the: 1,5; physical: 2; model: 3; of: 4; database: 6,10; is: 7; implemented: 8; using: 9; management: 11; systems: 12

Table 2. Reverse indexing table

in	file 1: 1,10; file 2: 7
many	file 1: 2
enterprises	file 1: 3
electronic	file 1: 4
data	file 1: 5; file 2: 5
is	file 1: 6; file 3: 7
stored	file 1: 7
and	file 1: 8; file 2: 10
processed	file 1: 9
computer	file 1: 11
systems	file 1: 12; file 3: 12
search	file 2: 1
engines	file 2: 2
work	file 2: 3
with	file 2: 4
stored	file 2: 6
the	file 2: 8; file 3: 1,5
database	file 2: 9; file 3: 6,10
entered	file 2: 11
parameters	file 2: 12
physical	file 3: 2
model	file 3: 3
of	file 3: 4
implemented	file 3: 8
using	file 3: 9
management	file 3: 11

The question of choosing keywords is important to prevent an increase in the number of words and to optimize the search. Connecting words and suffixes are not affected if they are removed from the text.

For example, in the above indexing tables, the core of the words “bases”, “base’s” and “the base” is the word “base”, and the rest of the words are formed by adding various suffixes. Just enter the word "base" in the indexing table. In addition, there are many unions in the text, such as “and,” “and,” and “also”.

Failure to participate in the indexing process of a combination of connecting words and possible suffixes for each word leads to a significant reduction in the index base and further search optimization. Many search engines have a dictionary with connecting words and affixes for English, Russian and other languages.

SEARCH FOR INFORMATION AND SORT RESULTS:

The search for information is based on a query entered by the user, and the list of files matching the query is a search result. The search process is performed in the next two steps:

- divide the request into keywords;
- Find keywords in the index database.

The division of the query into keywords is performed as described in the indexing process. That is, removing connective words and suffixes from the contents of the request [2]. The search for keywords in the index database is performed differently according to indexing methods. In the proper indexing method, keywords are searched from indexed words in files. If the keyword is present, the file is added to the list of results. For example, if the word “data” is searched in the above table 1, the search result will be 3 files, since the word is present in all three files.

In the reverse indexing method, the keyword is placed before the index table, and as a result, a list of relevant and previously created files is considered. For example, in table 2 the word “data” is found, and 3 files in the corresponding list are search results.

One of the most important steps of search engines is to sort the search results. During the sorting process, the results are sorted according to the degree of proximity to the user's request. Thus, the first file presented as a result is the file closest to the request, and the rest of the files are sent in the same order. Sorting the results is based on the number of occurrences and serial numbers of words in the index database. If the request consists of several phrases, the files in which each word is involved are identified and sorted according to how many of them are in the file, how close their serial numbers are to each other, that is, how they are arranged in sequence.

CONCLUSION:

The results of the analysis show that the reverse indexing method is effective when a search is performed on a large number of files in a search engine. For this reason, popular search engines use reverse indexing. But if the set of files in the electronic document database is in a certain sense limited, using the right indexing method will have the expected effect.

The article analyzed the working mechanisms of existing modern search engines. The results of the analysis can be used to develop general functional requirements for a database of electronic documents when creating an advanced search system.

Based on the foregoing, in further studies we will study the search algorithms used in modern search engines, and database indexing algorithms.

REFERENCES:

1) N. A. Gaydamakin. Automated information systems, databases and banks. M.: Gelios ARV, 2002. 259-260p.
2) A.V. Agranovsky, R.E. Harutyunyan. Indexing document arrays http://www.osp.ru/pcworld/2003/06/165855/
3) A.V. Kirillov. Features of Google Desktop Search. http://hostinfo.ru/favicon.ico
4) New Google product contributes to data leakage. http://www.securitylab.ru/favicon.ico
5) Personal Yandex search. http://pc4me.ru/personalnyiy-poisk-yandeksa.html
6) Sergey Koksharov. The principle of the search engine. https://devaka.ru/articles/how-search-engines-work.
7) What is Google Desktop? http://blog-vitalika.ru/?feed=rss2&p=269.
8) Search Algorithms in Arrays http://school-collection.lyceum62.ru/ecor/storage/3ab4a160-cf85-4876-b8da-9bbb1099ac8f/[INF10_04_12_TI_1C].html
9) Search Algorithms http://algol.adept-proekt.ru/algoritms_poiska
10) Snippet, search back algorithm, page indexing and working features of Yandex http://joomla-s.ru/interesnye-stati/51-seo/uchimsya-nravitsya-yandeksu-i-google/158-cnippet-algoritm-obratnogo-poiska-indeksatsiya-stranits-i-osobennosti-raboty-yandeksa.