COMPARATIVE ANALYSIS OF THE QUALITY OF DATA OBTAINED FROM INTERNET SEARCH ENGINES

Živković Miloš[1], Blagojević Marija[2], Stanišević Ilja[1]

[1] Western Serbia Academy of Applied Studies, department Valjevo, Serbia
[2] Faculty of Technical Science Čačak, Serbia

Abstract

Nowadays, the need for data is increasing, and in order to access data, we use a browser on a computer or mobile phone. Today's search engines give us a large number of pages for a search term. The question arises whether the search engine used for the search will really provide us with quality information about the requested data? In order to provide an answer, it is necessary to clarify how search engines get data from all the sites where the same is published, what do they do with it and how do they show it to users? The aim of this paper is to try to find out the quality of internet searches that we get from the most commonly used search engines.

Keywords: Web mining, internet search engines, search quality of internet data.

INTRODUCTION

The need for data is increasing, and in order to access data, we use a browser on a computer or mobile phone. Will the search engine used for the search really provide us with quality information about the requested data? Questions arise as to how search engines get data from all the sites where it is published, what do they do with it and how do they show it to users?

Today's internet data collection techniques include web mining. Data mining is a process that discovers significant relationships, patterns and behavioral trends in large amounts of data stored in warehouses that are accessed through recognition techniques, as well as statistical and mathematical techniques.

EXPOSITION

There are two approaches that define web mining. First is the "process-oriented view" (Etzioni 1996), which describes web mining as a series of jobs that need to be done. Another approach is the "data-centric view" which defines web mining according to the types of data used in the mining process itself [1].

The second definition is more acceptable today, and we can say that Web mining represents a set of applications that access data by researching the web, pages, their content, documents associated with those pages, but also by monitoring logs of access to those pages, time of use and other relevant data. which can be obtained when using them [2].

Data mining is the process of discovering meaningful new correlations, patterns and trends by looking at large amounts of data stored in various forms of patterns, using pattern recognition technologies and statistical and mathematical techniques.

Since the amount of data is very large, it is practically impossible for the user who wants to do some analysis of the same to do it independently without auxiliary tools. Therefore, it is necessary to have tools that summarize the data and provide the necessary views or information that could help analyze the data to improve the business.

TAXIOMETRY WEB MINING

According to the types of data being mined, web mining can be divided into three different categories as shown in Figure 1:

a) Content mining
b) Data structure mining
c) Mining using the Web [2]

This paper deals with the mining of the data structure and the content of the obtained data.

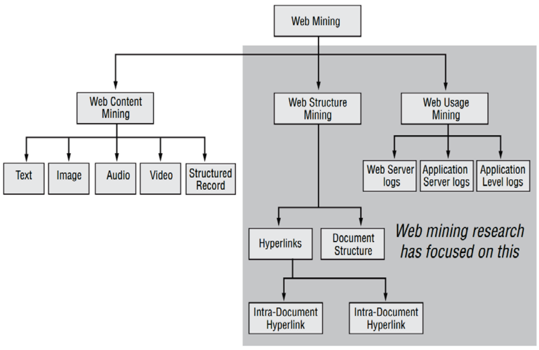


Fig. 1*: Web mining Taxonomy [1]*

WEBSITE CONTENT

The web page itself is an XML or HTML structure i.e., structured data that is displayed to the user. This structure contains tags i.e., marks that are important for displaying data, but also meta tags that, as a rule, should describe the information on the page in more detail.

Unfortunately, there is no standard for using meta tags, so we have various examples

<meta content="0;url=/search?q=fiscalization+satisfaction+blog&amp

<meta http-equiv="refresh" content="0;url=https://search.yahoo.com/search?p=fiskalizacija+zadvodstvo+blog

It's the same with other tags

<link href="/search?format=rss&amp;q=fiscalization+satisfaction+blog&amp;

href="https://r.search.yahoo.com/_ylt=AwrEoLyGr7xiaMsB_FpXNyoA;_ylu=Y29sb

<a href="/search?q=fiscalization+satisfaction+blog&amp

<a href="https://www.google.com/webhp?hl=en&amp;sa=X&amp;ved=0ahUKEwjY1f7

Each of our clicks on a link trigger some hidden activity on the page in order to display the requested page, but also to collect certain data about the user's behavior, so that the browser in the following period will try to fulfill our expectations as best as possible.

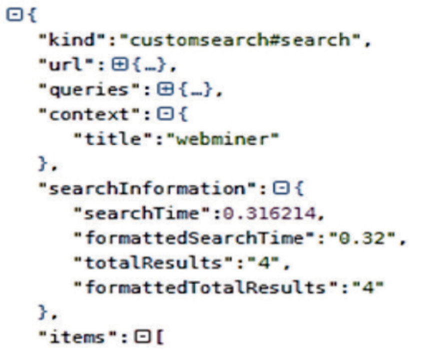
When the user opens a certain link, a page with the requested data can be displayed in front of him, but in addition to the requested data, that page also contains a lot of other data that we did not request, such as advertisements, but also data that the browser concluded that we need based on of our previous research.

OPERATION OF THE BROWSER

Search engines have their own search algorithms and methods. When searching, they read meta-tags, as well as the content of pages on the Internet. They do not analyze that data, but simply collect it, so that they can give an answer to the requested query. Whether those tags are written correctly or not, whether they really describe the term that is on that page or it was thrown in order to show that page to the user, the search engine does not know that.

Google is one of the first search engines to introduce the importance of link structure in mining information from the web. PageRank, which measures the importance of a page [1], is a core technology in all Google search products and uses the structural information of the web graph to return high-quality results.

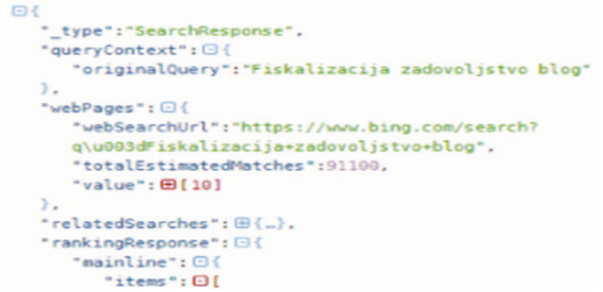
Google provides data through its API (Application Programming Interface) as shown in Figure 2.



```
⊟{
   "kind":"customsearch#search",
   "url":⊞{…},
   "queries":⊞{…},
   "context":⊟{
      "title":"webminer"
   },
   "searchInformation":⊟{
      "searchTime":0.316214,
      "formattedSearchTime":"0.32",
      "totalResults":"4",
      "formattedTotalResults":"4"
   },
   "items":⊟[
```

Fig. 2 *Data search results using the google API*

Microsoft, through its API, returns search results, but also gives us the ranking of each page based on its algorithms /rankingResponse/.



```
⊟{
   "_type":"SearchResponse",
   "queryContext":⊟{
      "originalQuery":"Fiskalizacija zadovoljstvo blog"
   },
   "webPages":⊟{
      "webSearchUrl":"https://www.bing.com/search?
      q\u003dFiskalizacija+zadovoljstvo+blog",
      "totalEstimatedMatches":91100,
      "value":⊞[10]
   },
   "relatedSearches":⊞{…},
   "rankingResponse":⊟{
      "mainline":⊟{
         "items":⊟[
```

Fig. 3 *Search result of Microsoft Bing search engine using API*

Every search engine that we turn to and through which we ask for the desired data shows us a page with a basic text that describes the requested data and a link that should open the requested page.

CONTENT OF THE LINK

The content of the link may take the user to a page describing the search term, but it may also take the user to a search engine page where the page describing the search term is embedded.

Anything to be displayed to the user, the user will search it for the required text.

Along with the requested text, the user will receive many other unnecessary data.

This paper does not deal with that excess data, but shows the quality of the data obtained from the search engine.

As mentioned above, content research is one of the aspects of web mining.

In order to explore the content of the page, it is necessary to display the desired page in the browser, by clicking on the desired link, and then press Ctrl+U, which will display the text view of the obtained page, where you can see the text view of the page, structured according to the xml rule /html structures. This text needs to be saved to a file and then analyzed.

The data structure starts like this:

```
<!DOCTYPE html>
<html lang="en-latn" dir="ltr" xml:
lang="en-latn"
xmlns="http://www.w3.org/1999/xhtml"
class=" responsive " style="">
```

The standard to be followed is given in the xmlns attribute [3].

In order to analyze such a text, it is necessary to have a program that will load all saved files according to the given criteria and display the search results.

If we look at the visual display of the page received from the search engine, it happens that the information we were looking for is not visible because it is not in the text we are reading, but in the meta tags or in the links that lead us to the following pages. Such pages generally do not meet the user's criteria.

However, on the pages we have links that direct us to pages that should have the requested content.

Opening the following pages - sublevels, recording the content and repeating the procedure is a demanding and complex task for the user.

Procedure for measuring the quality of search engine work

The procedure for quality assessment of the results obtained by the search engine includes the development of a program that has the following modules:

- establishing a connection to the search engine and setting a query for the required terms.
- analysis of the obtained data and opening of links to which the browser directs us,
- collecting data from the received pages and searching for terms and links on the received pages.

Since addressing search engines is not the same, it is necessary to create a program for each search engine. Some search engines do not allow you to contact them without prior registration and obtaining permission to access data, and some allow direct access, but the results are cluttered with advertisements and other unnecessary information.

The measurement technique applied in the project is finding the number of pages to which hyperlinks direct us, as well as the number of pages that contain the user's searched terms, i.e., data structure mining and content mining.

If we look at the pages we get through the search, and then the links on those pages that further direct us to the searched terms, we can measure not only how many of those hyperlinks there are, how many pages with the searched terms to which the links direct us - the depth of the pages, how many fake ones there are pages because they do not contain the required terms and hyperlinks led us to those pages, and measure the quality of the pages obtained.

If it is the number of pages (Np- total number of pages) that we received

$$Np = \sum_{i=1}^{n} p \text{..(1)}$$

and the number of pages containing at least one of the requested terms Ng (number of pages with text)

$$Ng = \sum_{i=1}^{g} p \text{ .... } (2)$$

we can say that the quality of the obtained results is in percentages

$$Q[\%] = \frac{Ng}{Np} * 100 \text{ .... } (3)$$

CONTENT SEARCH APPLICATION

Creating an application for page content search consists of two parts:

- Querying the search engine to obtain basic links to pages that have the required content;
- Search the resulting pages by depth as long as there is the required content.

The search of the obtained pages involves analyzing the content, finding the requested text, analyzing the data structure, searching for links to new pages within the obtained structure, and searching for the desired data.

When creating the application, it is necessary to take into account that some browsers do not allow you to contact them through the application if you have not previously registered with them, received certain codes and authorizations, that is, you must follow their rules, because for certain services you have to pay for using them.

In accordance with that, an application was created that allows search engines to be searched through the search engine API service and without the API service, by direct analysis of the data obtained from the search engine page itself.

API services give us data that is in the JSON standard [4], which we analyze and based on the criteria we get the addresses of the pages with the requested data.

When addressing the browser directly, we get a page that is in html format, where data about pages with our data is in meta tags with relative or absolute addressing.

The example search should exclude the search for commercial sites, and should only give us sites that describe the requested data. As there is a lot of talk about fiscalization in the period of writing this paper, we decided to search for sites that contain the following terms Fiscalization Pleasure Blog.

By querying the search engine through the application, we get a basic set of pages that have one or all of the terms. The page set is the same as if we manually set it up within the browser.

From the received response of the search engine, we extract the URL to the pages that, according to the search engine, should contain the requested data, and then we start analyzing the content and open all the links that further direct us to the requested data.

Because the number of URLs that appear on the pages, we investigate further is large, we had to limit the depth of the investigation to 30 levels and up to 30 sublevels.

Display of results obtained by content search and page structure search is displayed in Table 1.

Table 1 *Display of results obtained by content search and page structure search*

Search engine	Number of pages	Number of sublevels	Number of pages with the search term	Number of links	Number of links with the search term	Number of links without repetition
Google	2433	2060	204	2035	1454	372
Yahoo	56833	51876	1294	50592	19513	4956
Google – API	3384	3212	171	3210	3023	173
Bing - API	20208	19146	1001	19137	15528	1070

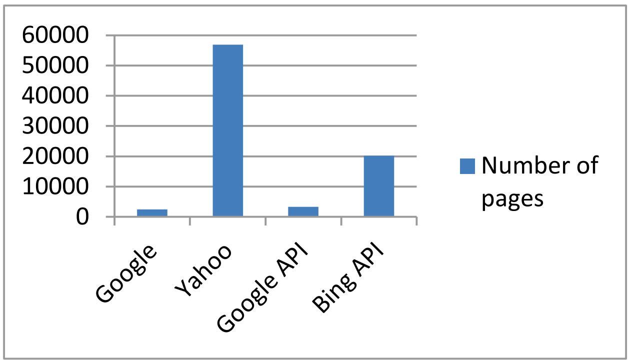


Fig. 5 *Display of the obtained search pages*

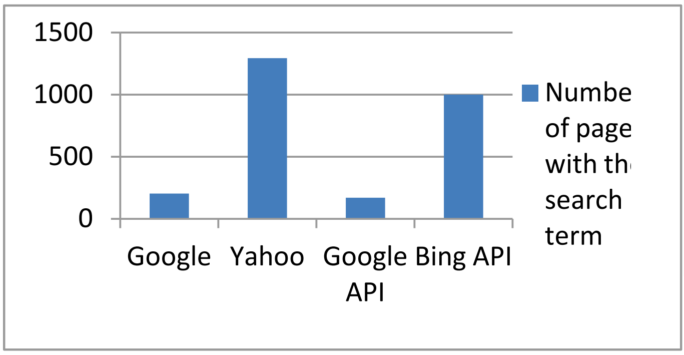


Fig. 6 *Display of the obtained number of pages with the search term*

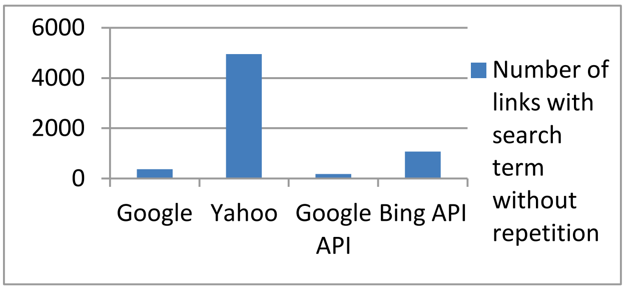


Fig. *7 Display of links pointing to content pages*

If the URL contains, for example, the text product/efiskalizacija-sa-cs-30-pro-na-1-klik/, that link is opened, because it contains one of the search words, and one of our terms is searched for within the text of that page.

The obtained results are divided into several files: the list level, sublevel, URL, requested text shows the linear structure of all the obtained content; list of URLs; A filtered list of URLs representing only non-repeating URLs.

Since the result can be both in Cyrillic and Latin, it was necessary to analyze the requested data in both letters, in order to properly evaluate the obtained results.

The results obtained differ from the initial results. Each search engine has its own criterion, as I mentioned at the beginning, and in addition to the criteria of page ranking, there is also the criterion of paid advertising, which will place the page in the first place.

If we look at these three graphs, we can see that the number of pages obtained from Yahoo and Bing search engines is much higher than the number of pages obtained from Google search, regardless of whether the search was done via API or not.

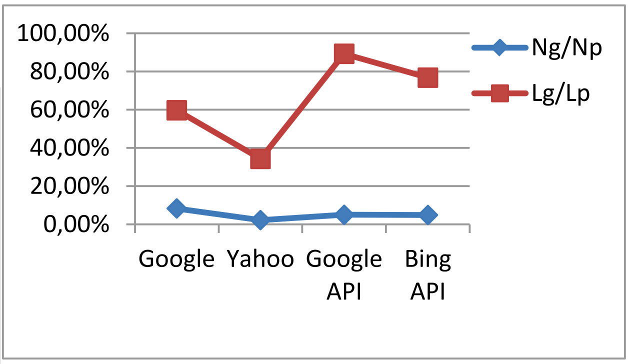


Fig. 8 *Ratio of received pages with the term and links with the term according to the total number of pages and links*

Figure 8 shows the ratio of received pages with one or more search terms to the total number of pages to which the links direct us.

On the part of the user who reads those pages, he may get the wrong picture of those pages because he will not find the terms, he is looking for on many of the pages he is directed to via links.

We also see that API search results are much better than search results, because search engines insert their links into user pages, to either place ads, or otherwise direct the user to an assumed search term.

CONCLUSION

Today's search engines give us a large number of pages for the searched term or terms.

If the page - presentation is well done, where the metatags and links correctly describe the content of the page itself, we will easily get to the requested data.

Through this work, we have tried to show how high-quality search engines are, through which we reach the required search results, dealing with the analysis of both the structure of the obtained data and the content of the obtained data.

We have seen that through the API methods of the search engines themselves, we can get not only better search results, but also other measurement data that the search engine itself has reached, and the same cannot be obtained inside the search engine itself.

It is also noticeable that many pages have metatags that are not adequate to the content of the page itself, which is probably the result of copying the content of previous pages without correcting the metatags themselves.

Further research could be done in checking the quality of the pages themselves by comparing the content of the meta tags and the content of the text that is displayed to the user.

REFERENCE

[1] Srivastava, Jaideep, Prasanna Kumar Desikan and Vipin Kumar. “Chapter 21 Web Mining — Concepts, Applications, and Research Directions.” (2004).

[2] Pranit B., Chawan P.M., "Web Usage Mining," Journal of Engineering, Computers & Applied Sciences (JEC&AS), vol. 2, pp. 34- 38, June 2013.

[3] Brahmia, Z.; Grandi, F.; Bouaziz, R., International Journal of Web Information Systems, Vol. 16, Issue 1, pp. 23-64.; Emerald Publishing Limited. September 02, 2019.

[4] Brahmia, Z., Grandi, F.; Brahmia, S., Bouaziz, R., Procedia Computer Science, 2021, pp. 184:823-828 Language: English. DOI: 10.1016/j.procs.2021.03.102.