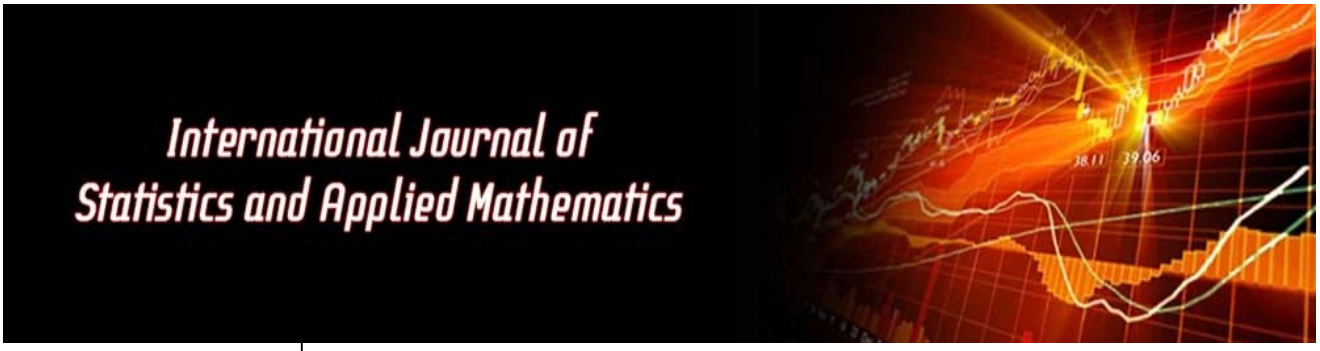

ISSN: 2456-1452
Maths 2020; 5(5): 30-33

www.mathsjournal.com
Received: 16-07-2020
Accepted: 20-08-2020

Prabhu Pant
Department of Information Technology, College of Technology, GBPUA&T, Pantnagar, Uttarakhand, India

Pankaj Joshi
Department of Information Technology, College of Technology, GBPUA&T, Pantnagar, Uttarakhand, India

Sanjay Joshi
Department of Information Technology, College of Technology, GBPUA&T, Pantnagar, Uttarakhand, India

Corresponding Author:
Prabhu Pant
Department of Information Technology, College of Technology, GBPUA&T, Pantnagar, Uttarakhand, India

A comparative study of search engines results using data mining and statistical analysis

Prabhu Pant, Pankaj Joshi and Sanjay Joshi

Abstract
Web search engines are keys to the immense treasure of information. Dependency on the search engines is increasing drastically for both personal and professional use. It has become essential for the users to understand the differences between the search engines in order to attain a higher satisfaction. There is a great assortment of search engines which offer various options to the web user. Thus, it is significant to evaluate and compare search engines in the quest of a single search engine that would satisfy all the needs of the user.

Keywords: Search engines, data analysis, URL

Introduction
Web search engines are among the most sought after tools over the internet. Millions of users access these search tools in quest of information from various spheres of life such as technology, tourism, travel, current affairs, literature, music, food, science and many more. Search engines have a huge database to which millions of pages are added everyday. Availability of pages searched by the search engines is dynamic, which means that the pages retrieved previously for a search query may not be available any longer as it might have been deleted by the author or turned obsolete. Web has grown to such an extent that it is not possible for a single search engine to crawl the entire web. Hence, many search engines are available over the internet which covers different portions of the web to select the results relevant to the search query which is further filtered and ranked before getting displayed to the user. Many search engines are available these days such as Google, AltaVista, Yahoo, Mamma, Infoseek, Lycos, Dogpile.com, MSN etcetera. The number of public web search engines available to the users is increasing to meet the growing size of the web and immensely increasing number of search engine users. Each search engine has a web database and the search results displayed to the user is a subset of the URLs contained in the database. Web searching is the second most popular activity following e-mail, as per Pew Internet study of web search engine users [7]. There is a significant requirement to identify the strengths of the search engines to utilize them efficiently for the desired results. Why a huge variety of search engines? How do these search engines vary? Which search engine suits my needs? How big the search engines are? Do they offer the same results? These are the questions that we have tried to answer in this paper with suitable methods devised by some researchers. The aim of this paper is to provide information about the design, working and results of search engines. This paper represents various evaluation methodologies to estimate the capabilities of search engines. This would help users to appreciate and select a search engine appropriate to their specific search needs. It may also facilitate the web search engine developers to suggest improvements in the web search engines. Section 3 describes the characteristics of the search engines in general focussing on the impact of these characteristics on the performance of the search engines. Section 4 illustrates the methods devised by various researchers to compare the search engines from different perspectives along with their limitations. Results of these methods have been elucidated in section 5. Summary & conclusion of the paper have been enlisted in section 6. Future work and acknowledgement have been specified in section 7 and 8 respectively.

Characteristics of search engines

There are countless web search engines available, still each engine offers a different result set to the users. This is because of the fact that there are certain features that make a search engine distinguishable from others. Some of the characteristics of Search Engines are as discussed below.

Web crawling or Spidering: A web search engine identifies the data available in the huge sphere of internet. Web crawler is a web robot that visits the list of URLs called seeds. The URLs identified by the web crawler are mined from various resources and are then downloaded to its own web database. Web crawling is carried over by web search engines recursively to offer up-to-date data to the users. There are some significant factors such as large volume of the data mined, high rate of modification, dynamic page generation, ever increasing web database and many more that may cause web spidering to pose challenges.

The behaviour of a web crawler is defined by a combination of the following policies [10]

- Selection policy: This policy defines the selection criteria for the pages that are to be downloaded to the web database of the search engine.
- Re-visit policy: It defines the frequency with which the crawlers check for updates and changes.
- Politeness policy: It defines the policy that would avoid clogging of the web sites.
- Parallelization policy: Interaction among distributed web crawlers is decided by this policy.

Result matching: It is a matching technique used by a search engine to match the user query with similar web pages existing in the web database. There are many different matching techniques employed by various web search engines to depict strongly relevant results. However, there can be challenges during matching of the results. Some of these are as discussed below:

- **Parsing:** Parsing algorithms may pose difficulties if they encounter complex Hyper Text Markup Language (HTML) used in some of the web pages. Such difficulties can create instances where some useful results may not be extracted for display to the user.
- **Filtering:** A search engine needs to perform effective filtering in order to show the most relevant URLs to the searchers. It is really significant to show unique results to the user by minimizing the chances of redundancy.

Result ranking: It defines the order in which search results are shown to the user. There can be thousands of results that may be shown to the user but showing results in order of relevancy needs to be taken care of. The best scenario would be when the user encounters results relevant for him/her on the first two pages. It helps the user to view the most relevant pages first. Search engines follow a sorting algorithm to rank the results. This algorithm counts on two factors:-

- **Location:** It is important for the search engine to look out for the search keywords at the top of a webpage. For example: looking for the search keywords in the title of a webpage.
- **Frequency:** The algorithm looks out for how frequently are the search keywords repeated in the context of the search results. Frequency of search keywords is not considered to be an ideal factor as it gets biased to content-rich pages [4].

Single-source search engines and Meta-search engines: Search engines can either have a single source or multiple sources of data. A search engine which extracts data from a single web database is termed as a single-source search engine whereas a search engine which extracts the most relevant results displayed by various single-source search engines is termed as meta-search engine.

A meta-search engine offers far more coverage than a single-source search engine. It facilitates the user to see the relevant data from many web databases at once using a single search operation. Examples of single-source search engines: Google, yahoo, AltaVista etc. Examples of meta-search engines: Dogpile.com, Mamma etc.

Web Indexing: After the web is crawled, search engines parse the document to generate an index that points to the corresponding result. The process involves concepts of mathematics and computer science for creation of indexes. These indexes help search engines for speedy retrieval of results. For example, it may take many seconds for a search engine to scan a document containing thousands of pages while on the other hand indexes can be really powerful for search engines as they eliminate the need for search engines to go through the document at the time search results are needed. Limitation of web indexing is that it needs extra memory for the storage of indexes. However, this limitation is compensated by the enhancement in performance that it offers [9].

Method

To compare the search results of different search engines and the uniqueness of each result, we can estimate the overlapping of search results and the relative size of search engines.

Search engines used in this study are - DuckDuckGo, Ask and Bing, mentioned in this study as E1, E2, and E3 respectively.

To collect the search result URLs of various search engines a python web scraping program is used, which stores the URLs of the first page of each search engine. This step estimates the coverage and intersection of the search results of search engines E1, E2, E3 and E4.

Data Collection- Select URLs from the results and store it in a array for further inspection

Result of the searches are stored in arrays named A1, A2, and A3 for search engines E1, E2, and E3 respectively. In order to have balanced queries we extracted 10 URLs of the front page from each search engine.

We used the results only from the first page because displaying appropriate results in the front page makes the search algorithms of the search engines efficient. Therefore, each search engine using their own algorithm will have different results and using this metric we can compare the results.

Comparison metric: To compare the results of the search engines, we used bit manipulation. The operators used with their purpose are below:

- AND - AND operator was used to compare the intersection of the search results
- OR - OR operator was used to find the union of all the search engine results

To find the difference of the results, we used the (-) operator.

All the three operations performed on the array has been represented below as venn diagrams of the sets.

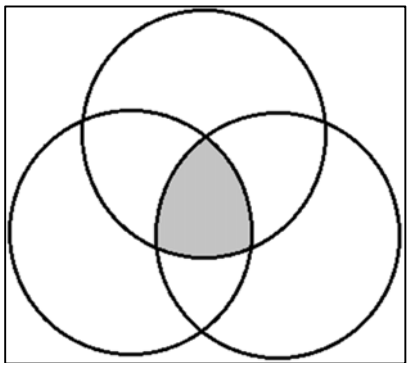

Fig 1: Intersection of sets

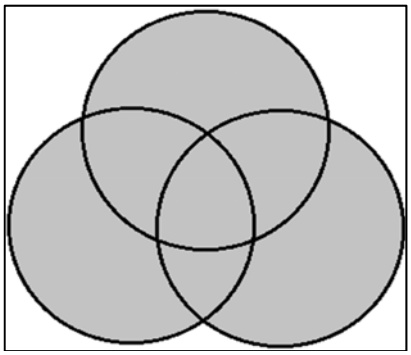

Fig 2: Union of sets

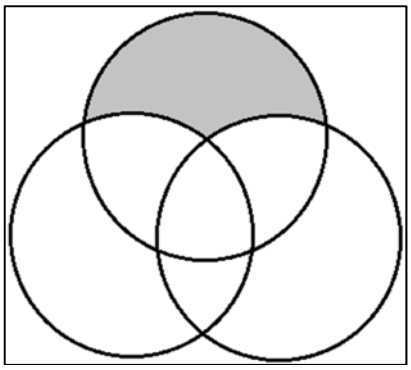

Fig 3: Difference of sets

Results

While formulating the results biases like query bias, ranking bias and statistical sampling error have not been considered.. Analyzing the search results, the links can be of the following types:

Dead link: A URL link is considered to be "dead" if it does not respond to the user's request. It may become dead due to any temporary problem or in case the page disappears during the time indexing was performed. A URL was graded '1' if it did not respond to the user's request otherwise it was graded '0'.Search engines were analysed for dead link in manual and automatic requests.

Topic: This feature tested various URLs if they were off the topic. The results showed that the proportion of such web links were particularly high for Yahoo that provided 49.1% off topic pages of the total count.

Commercial site: It may happen that search engine results redirect the user to some online sales and transaction site. The user may not be aware of these sponsored links. As per the observations Amazon, eBay and Flipkart were the most commonly returned commercial websites due to their popularity.

Relevance: It shows the degree to which the search results are related to the user's needs. Google and Bing produced most relevant results. The overlap keeps on decreasing as the number of search engines increases as depicted in Fig 4. The search engines which were used are (in order) - DuckDuckGo, Ask, Bing, Baidu and Parsijoo.

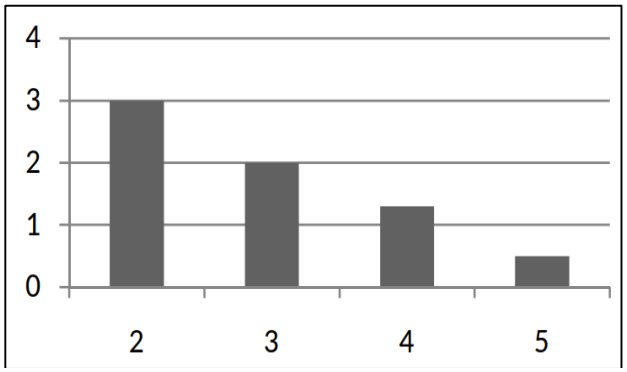

Fig 4: Degree of overlap between search results X -> number of search engines Y -> percentage of result overlap

The degree of correlation has been calculated using Karl Pearson's Degree of Correlation formula, which is as follows:

$$r = \frac{\sum xy - \bar{n}\,\bar{x}\,\bar{y}}{(n-1)\,SD(x)\,SD(y)}$$

Where
r = coefficient of correlation
n = number of items in each set
SD = standard deviation
x, y = items of sets X and Y respectively
$\bar{x}$, $\bar{y}$ = mean of x and y

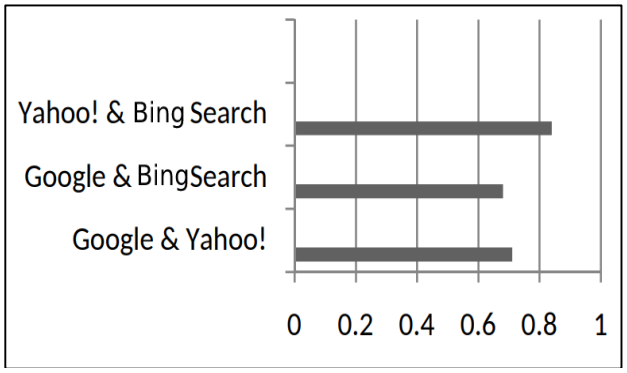

Fig 5: Degree of correlation. X -> Degree of correlation Y -> Web search engines

For calculating the precision of the search engine results, the following formula was used:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

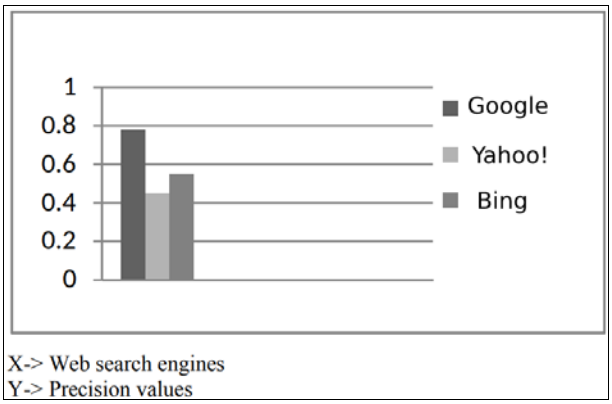

Fig 6: Search precision value

Conclusion

A web search engine opens the door to explore a huge amount of information. There is a variety of search engines which offer diversified services to it's users. This paper draws a clear picture of the differences between various search engines and disproves the notion that all web search engines have same search capability, coverage, ranking and indexing techniques. Web search engines differ from each other in multiple aspects such as the searching strategy, coverage of the web, relevance of the search results with respect to the search query, ranking of the search results etcetera. The overlapping of the search results offered by the search engines is very low. The overlapping of the results from various search engines could be measured by collecting sample URLs from the result set of a search engine for a specific query. URLs from the collected data can then be matched with the results of another engine by performing a string comparison. The number of matches could be recorded to determine the fraction of URL overlap. The results of the comparative study helps the users understand and select search engines that meet their needs. For example, search engine marketers get to know that a single search engine has a limited coverage so may not be sufficient for advertisements through sponsored links, thereby assisting the user to select many singlesource search engines or a meta-search engine for advertisements depending upon the need. Almost all search engines display the search result in the decreasing order of relevance to the search query except Voila which follows the reverse order. The specified results are not static as the web database keeps on changing, therefore, dependency on these results can not be forever and hence periodic search engines comparisons are advised.

References

1. Krishna Bharat, Andrei Broder. Estimating the Relative Size and Overlap of Public Web Search Engines, DIGITAL, Systems Research Centre, 130 Lytton Avenue, Palo Alto, CA 94301, U.S.A. 7th International World Wide Web Conference, 1998
2. Amanda Spink, Bernard J. Jansen, Vinish Kathuria, Sherry Koshman. Overlap among major web search engines. Emerald Group Publishing Limited. Internet Research, 2006, 16(4).
3. Heting Chu, Marilyn Rosenthal. Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology, Proceedings of ASIS 1996 Annual Conference: http://www.asis.org/annual96/ElectronicProceedings/chu.html
4. Mike Thelwall. Quantitative Comparison of Search Engine Results, School of Computing and Information Technology, 2008
5. Jean Véronis. A comparative study of six search engines. Université de Provence, Version 1.0 (en) – 22 février 2006: http://www.up.univ-mrs.fr/veronis, http://aixtal.blogspot.com.
6. Bernard J. Jansen, Amanda Spink. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management, 2006, 42(1).
7. Bernard J. Jansen, Amanda Spink, Chris Blakely, Sherry Koshman. A study of results overlap and uniqueness among major Web search engines. Information Processing & Management, 2006, 42(5).
8. http://internet.suite101.com/article.cfm/what_is_dogpile
9. http://en.wikipedia.org/wiki/Index_(search_engine)
10. http://en.wikipedia.org/wiki/Web_crawler