A Micro Analysis of Queries in Search Engines

ABSTRACT

The Internet has become one of the basic needs of people and it is a medium to vast knowledge. To access that knowledge, we use different search engines like Google, Yahoo, Bing, etc. These search engines use web crawlers to find the relevant information provided by the user as keywords in Search Queries. As there are millions of websites on the internet, extracting the most useful content for the user is the task of a Search Engine. Making one's content/website optimized, so that one can get a higher ranking in any Search Algorithm is what we call Search Engine Optimization. This paper describes the differences between the 5 leading and most widely used Search Engines based on 5 queries. The top 20 results from each query are picked and analyzed to find how many of them are third-party web pages. Also, it analyzes if location places any role in providing better results. Based on these and some other related analyses, the best effective search engine is concluded.

Keywords: Search engines, Search Engine Results Page (SERP), Web-Links, crawlers, third-party links.

INTRODUCTION

A search engine query is an information request that is submitted using a search engine. A search engine query is made each time a user enters a string of characters and pushes the "Enter" key.

The search engine matches results with the query using a string of characters that serve as keywords. On the Search Engine Results Page (SERP), these outcomes are shown in descending order of importance (according to the algorithm).

The terms that are used include:
• **Search filters:** The most popular filter attributes and values
• **Search queries:** The most frequent searches and the ones that produce no results
• **Search results:** The outcomes that users receive the most frequently for all searches
• **Clicks:** Track clicks by users and the click-through-rate (CTR) for a certain query.
• **Conversion rates:** The conversion rates of all search queries.

Search engines employ a query to get records from an index in descending order of relevance. Users anticipate human-like relevancy and machine-like speed when searching, spanning a wide range of possible results. All of these expectations must be met by search engines.
When someone uses a search engine, they not only receive information in answer to their search but also give the search engine information about the items they are looking for, which the search engine may use to assist other users. The search engine tries to link the two search terms together if the user conducts a subsequent search that is related to their first query. The search engine may use the data from those search query sequences if users conduct a series of queries on a certain idea.

The search engine may determine that the phrases are semantically connected to one another if many users conduct the first search, the same second search, or the same search inside a search session. The search engine might be able to assist users to find items online and offer better adverts in addition to search results if it is aware of the relationships between search queries. As a result, we can also assert that search engines do more than just provide us with filtered results; they also store our query information and attempt to connect and discern patterns among them. These search engines may therefore be considered as a genuine application of data science since they both search databases for data and use the search query provided by the user to create distinct data and do analysis on it.

Functionality of Search Engines

Web crawlers are used by search engines to index billions of pages. Crawlers, also referred to as spiders or bots, search the internet for new pages by clicking on links. After then, these pages are placed in an index from which search engines fetch requests.

The three main operations that search engines do are as follows:

1. **Crawling:** Search the Internet for content while reading the code and content of each URL you come across. The first search engine "crawls" the internet in search of pages to include in its database. Finding fresh books to add to that library requires crawling.

2. **Indexing:** Archive and arrange the information gathered through the crawling operation. Then, after being "indexed" or organized, the findings are added to their database. Putting your books in a precise order is called indexing.

3. **Ranking:** Provide the material that best responds to a searcher's query; as a result, results are ranked from most to least pertinent.

Webpage Ranking

The main motive to rank a page is that it will help decide the search engine which webpage should appear at which position, meaning which webpage is more relevant when compared to the others. It will help in showcasing the results in the decreasing order of relevance to the query made.

The basic working of this process is that it first filters out all the relevant web pages based on the keywords found in the title of the webpage by matching it to the given query and then ranking these web pages on certain criteria. The criteria may differ for different search engines, but the base idea is that the webpage that is referred, the highest number of times and the percent of chances of that webpage being referred by others is highest, that webpage can be thought of to be more relevant. As such, that webpage will be given the highest rank followed by the next favorable webpage, and so on.

This algorithm of ranking is very crucial for deciding the efficiency of the search engine because it is directly responsible to show the results to the user. If the algorithm is not good, the user will have to scroll down a lot of SERPs to reach the required webpage.

PROPOSED METHODS

Firstly, we selected 5 topics which we used as our queries to be searched on the search engines.

The queries are:

1) IT companies in Bangalore

2) IT companies in Gurgaon

3) Colleges in Vellore

4) Colleges in Delhi

5) Universities in USA

Then, we ran these queries in 5 different search engines namely, Google, Yahoo, Bing, Ask, Ecosia, and found the corresponding results. From these results, we scraped the top 20 titles and the corresponding links of the websites appearing on the search engine result page (SERP) using selenium and selenium-pro package of python. These packages are meant for automation and thus the links and titles got scraped by the python programming automatically.

These links and titles from different search engines were then stored in python's pandas DataFrame, forming 5 different dataframes. Each DataFrame included links and titles of all the 5 queries from one search engine. We, then, used simple analytical techniques to analyze how many third-party links appeared in a particular query results and compared it over different search engines' results. Here, the third-party links mean the webpages that are not the direct webpage of the organization, institution or related body corresponding to the search query, but which are some other webpages that direct you to such original webpages. On the basis of these differences found in number of 3rd party webpages and original webpages, we made our conclusion as to which search engine is more effective and providing better number of original webpages.

Tools Used: Python (pandas package for storing Data and selenium, selenium_pro packages for scraping web-links and titles), Excel for making Tables.

RESULTS AND DISCUSSIONS

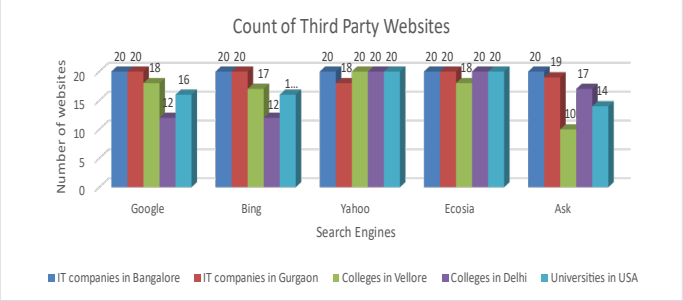
The number of 3rd party websites corresponding to different search engines are:

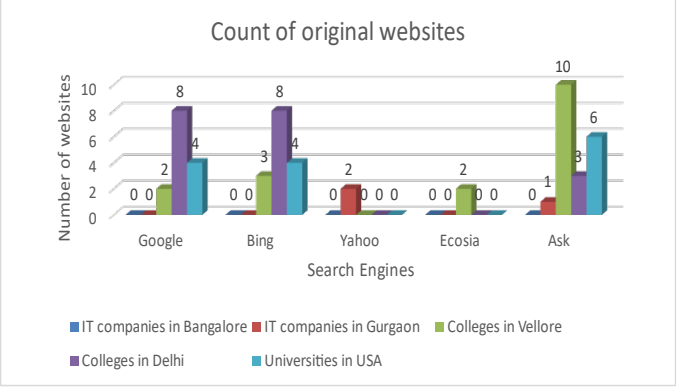
Count of 3rd Party Websites					
QUERIES	Google	Bing	Yahoo	Ecosia	Ask
IT companies in Bangalore	20	20	20	20	20
IT companies in Gurgaon	20	20	18	20	19
Colleges in Vellore	18	17	20	18	10
Colleges in Delhi	12	12	20	20	17
Universities in USA	16	16	20	20	14

The number of original websites corresponding to different search engines are:

Count of Original Websites					
QUERIES	Google	Bing	Yahoo	Ecosia	Ask
IT companies in Bangalore	0	0	0	0	0
IT companies in Gurgaon	0	0	2	0	1
Colleges in Vellore	2	3	0	2	10
Colleges in Delhi	8	8	0	0	3
Universities in USA	4	4	0	0	6

Graphical Representation of Number of third-party websites:





working in Bangalore and rather gave all the third-party websites. Similarly, when “IT companies in Gurgaon” was searched only Yahoo could give 2, and Ask could give 1 actual website of IT companies working in Gurgaon whereas the other websites were all 3rd party websites. However, the other 3 queries showed up mixed results in search engines other than Yahoo.

These differences can be seen because of the keywords in the query search. When “IT companies” was typed, the search engines focused on these words rather than seeing whether it was a website of any actual company located in Bangalore or Gurgaon.

Also, when colleges were searched, the results were better in the case of Delhi as compared to that located in Vellore. This can be because of the fact that there are more colleges in Delhi and fewer in Vellore. Similarly, the results were somehow better than others when foreign universities were searched. So, we can say that location might affect the ranking of the web pages too.

CONCLUSIONS AND FUTURE SCOPE

So, from the results obtained, it is clear that none of the search engines could give significant number of original websites on different queries performed. (Before reaching to any further conclusion it is important to tell in advance that results while searching “Colleges in Vellore” on “Ask” search engine, the results on SERPs were very few due to which it is giving false impression of performance of Ask in that field). However, from these results, we can say that Google, Bing and Ask are performing better than others. But keeping the limitation faced in Ask search engine in mind, we could conclude that in terms of providing the original webpages, Google and Bing are outperforming the other search engines. However, in terms of other aspects, different search engines might be more effective.

One could further work on including queries in different languages and then noticing how different search engines react to it.

Also, one could include more recent search engines which are providing better results but are, however, more protected when compared to the traditional search engines.

Further one could think of writing a more accurate page-ranking algorithm so that the search engines could involve looking for whether the webpages that it is displaying are third-party or not and ranking the 3rd party webpages accordingly by giving higher preference to original webpages.

We could not include recently developed search engines that are claiming better results from the traditional search engines like Brave, OneSearch, etc. The major reason was the privacy aspects that these websites are providing due to which the web scraping of these websites was not completed successfully. Also, there are other search engines like Chacha which are majorly meant for working in a particular part of the world. Such search engines could not be included due to location and language factors.

Also, the web results were very few when ‘Colleges in Vellore’ was scraped from the ‘ASK’ search engine due to which more number of original webpages came up and thus, leading to some false interpretations if that would not have been taken care of while analyzing.