A Survey on Search Engines

Shiwani Gupta
Assistant Professor
CMPN, TCET, Mumbai, India

Abstract

Web surfing for various purposes has become a habit of humans. Searching for information from the Internet today has been made easier by the widely available search engines. However, there are many search engines and their number is increasing. It is of considerable importance for the designer to develop quality search engines and for the users to select the most appropriate ones for their use. The Information quality linked through these searches is quite irregular. There are fair chances that the retrieved results are irreverent and belong to an unreliable source. In fact, most search engines are developed mainly for better technical performance and there could be a lack of quality attributes from the customers’ perspective. In this paper, we first provide a brief review of the most commonly used search engines, with the focus on existing comparative studies of the search engines. The paper also includes a survey conducted of 137 respondents where the identified user expectations will be of great help not only to the designers for improving the search engines, but also to the users for selecting suitable ones. The objective behind this study was also to find the reason behind poor precision and recall of so many available search engines. The study finally aims to enhance user search experience.

Keywords: Search Engine, Information Retrieval, World Wide Web

I. INTRODUCTION TO SEARCH ENGINES

Search Engine is a computer program that searches documents on World Wide Web for specified keywords. Crawlers are computer robots which actually build search engine databases by browsing through the internet/web to find pages which are potentially capable of containing results as asked for, and are present within these search engine databases. The drawback of these is that if any page is not linked to any other page via a link, then it’s not possible for the spiders to find it. The solution to this is, to put that brand new page as a link to already present pages or to add its URL manually for inclusion. This feature is already incorporated into every major search engine available online.

As soon as these web pages come into contact of any of these crawlers, another computer program, spider is on to its work for "indexing." The Indexing program is responsible for identifying links, texts, images and other content available in the web page and storing this page into the search engine database's files. Indexing these pages saves us from searching the whole web for the exact search keyword thus limit rework and saving time.

Such web pages which are not accessible by search engine spiders are excluded from the searchable databases uploaded on the web and hence such contents are termed as Invisible Web or Deep Web or Deep Net. Some search engines for Deep Web are Infomine, Intute, Infoplease, IncyWincy, etc.

Searching on the WWW for articles, theses, books, and abstracts from academic publishers, professional societies, online repositories, universities and other websites can be a lot easier with a focused search via Academic search engines like googlescholar, citeseerX, getCITED etc.

The same query on multiple search engines would return different results. Thus there are search engines called as Metasearch Engines which don’t maintain their own database but forward search terms to a number of different search engines and collate. Examples of such metasearch engines are DogPile, Metacrrawler etc.

The searchers today are quite keen to get quick results without the need to key in the keywords. Hence few Multimedia Search engines which can search Image, Audio and Video are Radio-locator, Blinkx, Pixsy, Retrievr, Picsearch etc. We all agree that we live in a Social Connect Era. Wherein, a business owner might want to tap into social search engines like Wink, iSearch, Pip, SocialSearcher etc. to get key information to customers.

“A progression from ubiquitous keywords to increasingly important entities”, words have become concepts and search engines have evolved to genuine Learning machines. Thus Semantic Search Engines seek to improve search accuracy by understanding searcher intent and contextual meaning of terms. To name few Kngine, Swoogle, SenseBot, Cluuz, Evri etc. retrieve results looking for sense on the web.

Moreover in recent years, search engines have increasingly pushed to monetize their organic search results, automate search query suggestion, evolve the search interface, place more focus on search personalization and localization and have built more advanced anti spam penalties and filters.
II. LITERATURE SURVEY

Web search engines can differ from one another in three ways – crawling reach, frequency of updates, and relevancy analysis. Therefore, the performance capabilities and limitations of web search engines, and the differences between single and meta-search engines, is an important and significant research area. There is a critical need for a greater understanding of the differences in web search engines’ web site indexing and the overlap among results for the same queries. Overall, most web searchers view only the first or second page of results (Spink and Jansen, 2004).

In [5], the author concludes that the percent of total results retrieved by only one of the three major web search engines Google, Bing, Yahoo was 85 percent, retrieved by two web search engines was 12 percent, and retrieved by all three web search engines was 3 percent. This small level of overlap reflects major differences in web search engines retrieval and ranking results. The findings point to the value of meta-search engines such as Dogpile.com in web retrieval to overcome the biases of single search engines. Further findings state that first page results returned by the three major web search engines included in this study are different from one another. Therefore, examining overlap levels for queries on first page results is an important research issue.

In [2] author concludes that international Search Engines like google.com, altavista.com and yahoo.com received positive comments in comparison to Greek search engines in terms of working environment, relevancy of data retrieved, satisfaction with precision, classification of results and quality of results, specially with Google.com.

In [3] author states that China, which has the largest Internet population and potentially the largest Internet market in the world, Google is far behind the local search market leader Baidu which holds around 70% of the search market while Google has only around 21%. Similarly in Korea, the local market leader Naver has 62% search market share and Google only has 4% (Bonfils, 2010; Market The- Globe, 2010; Visual Economics, 2010) because Google’s bare bone design of home page with a single search box is very popular in Western countries, but in Korea, people like web pages with rich contents so users can find a lot of information on one page easily.

In [1], the author has studied Google, GoogleChina and Baidu. He asserts that Google outperforms Google China in terms of retrieval effectiveness in five search features (title search, basic search, exact phrase search, PDF search, and URL search). Meanwhile Google China surpassed Baidu in terms of retrieval effectiveness in four search features (title search, basic search, PDF search, and URL search). The regression analysis results show that Google also exceeded Google China and Baidu in terms of search results ranking quality in all five search features. Google China outperformed Baidu and held the second place in this category. Google search engines outperformed the Baidu search engine due to factors such as quality of indexing in the databases, coverage of the databases, quality of webpage contents, quality of results rankings, and so on.

In [6], the author states that none of the search engines received more than 42% recall. Metasearch engines don’t give better results than individual. Searching several search engines is necessary to achieve complete search but metasearch engines are not a substitute.

In [4] the author summarizes that engines with multiple language support features were EZ2Find in the metadata search engine category, Google in the regular search engine category, and Onlinelink in the visualization search engine category. Google supported 37 different languages. Google’s translation tool was the strongest one of all the search engines. Secondly, the maximum number of languages a search engine could support in this investigation was 49 and the average number of languages supported per search engine was about 21.57. In the investigated search engines, the maximum number of language translation pairs was 18. Only three of the investigated search engines (Google, EZ2Find, and AltaVista) provided users with a translation feature.

III. EVALUATING SEARCH

The research question is whether search engines retrieve relevant documents or are they ranked high enough to be seen by user.

Following is the bases for statistical comparisons of retrieval effectiveness:

- If engine A retrieves ‘a’ relevant documents in top 20 and there are ‘c’ possible relevant documents, then
  - Precision=a/20 and Recall=a/c
- If engine B retrieves ‘b’ relevant documents in top 20 and there are ‘c’ possible relevant documents, then
  - Precision=b/20 and Recall=b/c

Therefore

- Precision/Recall=a/b.

![Fig. 1: Recall and Precision](image-url)
Thus we define Recall as how well a system retrieves all relevant documents whereas Precision as how well the system retrieves only the relevant documents.

Looking into the second research question, the author has identified the ranking algorithm used into some of the traditional search engines.

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Ranking algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>inbound links, authority, relevancy etc.</td>
</tr>
<tr>
<td>Baidu</td>
<td>no. of backlinks pointing to you site</td>
</tr>
<tr>
<td>Altavista</td>
<td>search term in title, in meta tag, beginning in body, exact phrase, etc.</td>
</tr>
<tr>
<td>Naver</td>
<td>updates search algorithm every year eg. Libra, sonar.</td>
</tr>
<tr>
<td>Yahoo</td>
<td>keyword in title, click popularity</td>
</tr>
<tr>
<td>Bing</td>
<td>click through rate</td>
</tr>
<tr>
<td>Yandex</td>
<td>links</td>
</tr>
</tbody>
</table>

### IV. COMPARISON OF SEARCH ENGINES

The author has used web analytics from [12] to find global ranking of few search engines, the visitor countries and major keywords searched.

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Traffic rank globally</th>
<th>Visitor Countries</th>
<th>Search keywords sending traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>google.com</td>
<td>1</td>
<td>U.S., India, China, Japan, Iran</td>
<td>Gmail, google, translate, maps</td>
</tr>
<tr>
<td>bing.com</td>
<td>30</td>
<td>U.S., China, Germany, India, U.K.</td>
<td>Weather, google, youtube, walmart</td>
</tr>
<tr>
<td>Yahoo.com</td>
<td>5</td>
<td>U.S., India, Taiwan, Indonesia, Brazil</td>
<td>Yahoo, youtube, google</td>
</tr>
<tr>
<td>Baidu.com</td>
<td>4</td>
<td>China, Japan, U.S.</td>
<td>Instagram, linkedin, java, watsapp, google</td>
</tr>
</tbody>
</table>

The author has also compared 10 search engines and found that though Bing is the latest in category but has been able to bag a position in the market. It was also found that Google and Bing are good at personalization of results but some searchers may want privacy which is provided by DuckduckGo. The table below enlists few search engines with their year of establishment, the country where found and the current status.

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Year Of Establishment</th>
<th>Country</th>
<th>Current Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>AllTheWeb</td>
<td>1999</td>
<td>Norway</td>
<td>Acquired by Overture in 2003</td>
</tr>
<tr>
<td>Altavista</td>
<td>1995</td>
<td>California</td>
<td>Lost to Google and purchased by Yahoo in 2003</td>
</tr>
<tr>
<td>Archie</td>
<td>1990</td>
<td>Montreal</td>
<td>First Internet SE, project was ceased in late 1990s</td>
</tr>
<tr>
<td>AskJeeves</td>
<td>1996</td>
<td>California</td>
<td>6th largest in U.S., Now known as Ask.com</td>
</tr>
<tr>
<td>Baidu.com</td>
<td>2000</td>
<td>China</td>
<td>Largest in China</td>
</tr>
<tr>
<td>Bing</td>
<td>2009</td>
<td>Washington</td>
<td>2nd largest in U.S. with query vol of 20.9%</td>
</tr>
<tr>
<td>DuckduckGo</td>
<td>2008</td>
<td>Pennsylvania</td>
<td>Emphasizes protecting searcher’s privacy</td>
</tr>
<tr>
<td>Google</td>
<td>1998</td>
<td>California</td>
<td>largest in U.S. with query vol of 63.9%</td>
</tr>
<tr>
<td>Infoseek</td>
<td>1995</td>
<td>California</td>
<td>Originally hoped to charge for searching</td>
</tr>
<tr>
<td>Inktomi</td>
<td>1996</td>
<td>California</td>
<td>Aquired by Yahoo</td>
</tr>
<tr>
<td>Lycos</td>
<td>1994</td>
<td>Massachusetts</td>
<td>NoheWebw outsources for its search results from AllIT</td>
</tr>
<tr>
<td>Yahoo</td>
<td>1995</td>
<td>California</td>
<td>3rd largest in U.S. with query vol of 12.5%, largest in Japan</td>
</tr>
<tr>
<td>Yandex</td>
<td>1997</td>
<td>Moscow</td>
<td>4th largest in Russia</td>
</tr>
</tbody>
</table>

The author knows that the ranking algorithms used also play a major role in retrieving relevant results and hence has analyzed the ranking algorithm behind major Search Engines, though much has not been talked about in literature. Google uses Inbound Links, Authority and Relevancy measures mainly alongwith with many others. Bing uses Click Through Rate whereas

### V. SURVEY

Users need to understand web search engine capabilities, coverage and limitations. Single web search engines have obvious strengths and weaknesses. In some circumstances, the uniqueness of a web search engine’s coverage may be useful for engine users. Our study also has implications for web search engines users. However, this information is not easy for web users to find.

The objective behind survey on search engines was to gather information that:
- Which search engine is most favored one
- Which search engine empowers which nation
- What are liked features of a search engine
- Which feature the respondents admire in some other search engine
- Lastly which feature of the search engine they don’t use
VI. RESULTS

The survey of 137 respondents of which 135 were Indians shows that the respondents were mainly using Google search engine.

The respondents have further come up with following good feature listings of Google:

A. Features of Google liked by users
   - Clean UI
   - Simplicity
   - ImageSearch Pro
   - Quick Response Time
   - Relevant/efficient results
   - Auto complete/correct
   - Smart prediction/suggestion
   - Search results pertain to search habits of the person querying
   - Is able to synchronize information on all devices
   - Use of operators for relevant search
   - Vernacular support
   - Google Earth
   - Google Maps
   - Google Scholar

They have also come up with some feature listings which they don’t use:

B. Features of Google not used by users
   - Voice search
   - Video search
   - “I’m feeling lucky”
   - Google+
   - Results after page 2/3
   - Google Wallet
   - Special Operators
   - Multi Language Support

Lastly the survey stated some features in some search engines which were present in other engines and not in the one they used:

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Feature not present/inappropriate</th>
<th>Present in which search engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>News Yahoo</td>
<td>Yahoo</td>
</tr>
<tr>
<td></td>
<td>Semantically connected results</td>
<td>Semantic search engines</td>
</tr>
<tr>
<td></td>
<td>Video play on hover Bing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deep Search Duck duck go</td>
<td>Deep web search engines</td>
</tr>
<tr>
<td></td>
<td>Privacy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Too many Advertisements Duck duck go</td>
<td></td>
</tr>
</tbody>
</table>

VII. CONCLUSION

The respondents of our survey were mainly Indians who claim that Google is the most sought after search engine platform. Moreover I can comment that every search engine has its PROs and CONs. There are few search engines, few features of which the users are not aware. Designers of search engines focusing in a specific country should keep in mind the needs of local natives.
for better acceptability. The survey can be of benefit both to designers of search engines to help them optimize their engine and to the users to decide which to go for which need. The paper summarizes that there is a lot of work done in the Search Engine Optimization field in past but there is still lot scope for research to happen.

REFERENCES

[8] www.googleinsidearch.com
[12] www.cse.google.com