Web Search Engines - A Comparative Study

Inderjeet Singh Oberoi
Mälardalen University
School of Innovation, Design and Engineering
Högskoleplan 1, Västerås, Sweden
ish09001@student.mdh.se

Mridul Chopra
Mälardalen University
School of Innovation, Design and Engineering
Högskoleplan 1, Västerås, Sweden
mca09002@student.mdh.se

1. 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. We have enlisted various methods to compare web search engines, along with their results and limitations to understand the commonalities and differences between the search engines. Web search engines have very less overlapping in their search results and it is because of the different indexing, ranking and crawling techniques used by them. This paper helps the users to find a suitable search engine for them and comprehend the basics of web search engines.

Keywords
URL (Uniform Resource Locator), sponsored links, non-sponsored links, HCE (High Count Estimate), ranking, sampling.

2. INTRODUCTION
Web search engines are among the most sought 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. All the references used for writing this paper have been enlisted in section 9.

3. 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.

1. 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.

2. **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 [4].

- **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.

3. **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 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]

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 etcetera.

**Examples of meta-search engines**: Dogpile.com, Mamma etcetera.

5. **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].

### 4. METHODS

Following are the methods to compare web search engines:

#### 4.1 Overlapping and uniqueness of web search results

Following are the two methods to measure the degree of overlapping and uniqueness of web search results:

**4.1.1 Method One**

**Purpose** - To estimate the overlapping of search results and the relative size of search engines

**Web Search Engines** - AltaVista, Excite, HotBot, Infoseek

**Devised by** - Krishna Bharat and Andrei Broder [1]

Collection of URLs indexed by the search engines followed by verification of the presence of each URL in the result set of specific search engines forms the basis of this method. This method helps in the estimation of coverage and intersection of the result sets of search engines $E_1$ and $E_2$.

**Data Collection** - Select URLs from the index of a search engine randomly using a procedure.

- **Procedure to collect data** - Create a lexicon consisting of the most frequently searched words over the web. Form queries using the words in the lexicon. In order to have balanced queries, arrange the words in the lexicon in decreasing order of the frequency by which they are searched. Use words that are equidistant from the middle of the lexicon to form queries.

There is a huge difference in the result set from two queries with same search words but different operators (conjunction (AND) and disjunction (OR)). Therefore, the set of queries to be used for data collection should be a
blend of queries with conjunction operator and queries with disjunction operator. Finally, execute this set of queries on various search engines and collect samples of URLs from the first result page randomly.

Verification - Verify if a particular URL from the collected data is contained in the result set of a specific search engine.

- Procedure to verify presence of a sample URL in the result set of a specific search engine- A strong query to uniquely identify the desired page should be used for the search page under examination. A query is considered to be strong if it fetches the page under test.

![Types of URL comparisons](image)

Consider two single-source web search engines as $E_1$ and $E_2$. Where, $E_1$ is relatively bigger than $E_2$.

![Overlapping of Search engine results](image)

**Relative size** - It is hard to estimate the exact size of search engines but their relative size can be estimated using the following formula:

\[
\text{Size}(E_2) = (E_1 \cap E_2) \text{ with URL collected from } E_2
\]

\[
\text{Size}(E_1) = (E_1 \cap E_2) \text{ with URL collected from } E_1
\]

\[
\text{Result overlap} - \text{The intersection of the result sets from two search engines } E_1 \text{ and } E_2 \text{ is given by:}
\]

\[
\text{Fraction of URLs collected from } E_1 \text{ found in } E_2
\]

**Our advice** - A better estimate of the result overlap can be performed by determining the “fraction of URLs from the relatively small search engine found in the relatively large search engine”. It aids to determine the closest estimate of actual intersection of the results from search engines $E_1$ and $E_2$ as sample URLs collected from relatively small search engine has high probability of overlap with the result set of relatively large search engines provided the result set of the two search engines is not disjoint.

**Limitations**

Various biases such as ranking bias, statistical sampling error, query bias had not been catered.

- Query bias - Content-rich pages tend to respond more to queries.
- Ranking bias - Low ranked pages have low probability to get sampled.
- Statistical sampling error - The statistical analysis result varies with sampled data.

### 4.1.2 Method Two

**Purpose** - This method suggests a way to find the overlap and uniqueness between the first page results of single-source search engines and measure the extent to which a meta-search engine provides the result of highly ranked searches by single-source web search engines.

Web Search Engines - Google, Yahoo, Ask.com, Dogpile.com


**Steps involved:**

i. **Data collection** - The server access log files of a meta-search engine were used to enlist queries to be executed for data collection. Queries from a diverse set of users were ensured by selecting the queries on a week day and a weekend. Duplicity of queries was removed by eliminating redundant queries from the list. Different queries from the query list were executed in a sequence for various single-source web search engines and the first result page was stored in the database.

ii. **Data analysis** - An algorithm to match URLs on different search engines was designed and implemented for each URL in the result set for different queries. When a URL for a query on a search engine matched the URL for a query on one or more other search engines a duplicate match was recorded for that query. A statistical analysis was then performed on the overlap result of first result page search. A complete result set was formed from an empty set by adding the unique URLs from the current result set into the complete result set. The uniqueness of the URLs added to the complete result set was ensured by checking the
presence of the selected URL in the complete result set using “string comparison” method as discussed below.

String comparison- A simple comparison of the URL in the current result set with the URLs in the complete result set is used to determine whether the result is in the complete result set.

Limitations
Following are the two major limitations for the above specified method:
i. Multiple languages were not taken into consideration when collecting data.
ii. Exact string comparison used for matching URLs in the result sets of search engines is not a good method to find result overlap as it may neglect overlapping of two URLs (one with host name and the other with an IP address) with same page.

Data collection was restricted to first result page which reduces the scope of search result overlap estimation. Due to different ranking techniques used by search engines, the results on the first result page of a search page may be available on the second result page of the other engine thus limiting overlapping result.

4.2 Performance, Search capability and consistency of search engines
Following are the three methods to measure the performance, search capability and consistency of search engines:

4.2.1 Method Three
Purpose- To compare search engines on the basis of search power, performance, composition of web indexes etcetera.

Web Search Engines- Alta Vista, Excite, Lycos
Devised by- Heting Chu & Marilyn Rosenthal [3]

In this methodology three search engines namely Alta Vista, Excite and Lycos were chosen and compared on the basis of certain features such as search capabilities, retrieval performance, output option etc. The search engines were tested using ten reference questions that were chosen by librarians at Long Island University. These questions involved variety of single words, phrases, Boolean logic, general themes, specific topics and concepts. The Reference questions were modified in accordance with the corresponding syntax of each search engine to generate a search query.

Example
Reference Question: computers and learning disabilities
Search Query:
Alta Vista: computers +"learning disabilit*"
Excite: computers learning disabilities
Lycos: computers learning disabilities

The test environment consisted of two web browsers namely Lynx and Netscape. This method carried out evaluation on first ten web records for each search engine.

Evaluation Criteria
Web search engines were evaluated based on the following characteristics:
- Composition of web indexes
- Search capability
- Retrieval performance
- Output option
- User efforts
- Limitations

Limitations
This method analysed only first 10 web records for each search query, the result might have varied if a bigger result set would have been chosen.

4.2.2 Method Four
Purpose- To measure the consistency and international spread of results returned by search engines

Web Search Engines- Google, Yahoo, Live Search
Devised by- Mike Thelwall [4]

In this methodology of evaluation, webometric investigations were performed. These investigations aimed at determining the number of pages that match a query and then measuring the international spread of those pages. Three Web search engines namely Google, Yahoo and Live search were chosen to perform evaluation. This methodology was not performed to compare these search engines but rather to measure the consistency of the results returned by these engines.

The evaluation was performed using a data set of 1,587 words. These words were provided as input to Application Programming Interfaces (APIs) of the search engines specified above. The results were recorded for each query for the corresponding search engine in the form of total number of results (HCE) and URLs returned. A total of 1000 URLs were recorded for every test case.

Evaluation Criteria
Evaluation Criteria of this methodology depends upon following two factors
- Hit Count Estimates (HCE)
- Complete list of URLs

Limitations
The search terms used consisted of alphanumeric characters while results may vary for non-ASCII languages such as Arabic and Japanese. There is also a possibility that different search options such as advance search, phrase search may respond in a different manner and provide different results.
4.2.3 Method Five

**Purpose**- The purpose behind executing this methodology was to determine comparative data relating to search in a language other than English such as French.

**Web Search Engines**- Google, Yahoo, MSN, Exalead, Voila, Dir.com

*Devised by*- Jean Véronis [5]

This study involved comparison of search engines such as Google, Yahoo, and MSN along with three French search engines namely Exalead, Voila and Dir.com. This study performed its evaluation from the user’s perspective rather than experts evaluating search engines. The evaluation methodology was held in December, 2005 with the assistance of 14 first year students from University of Provence, who performed the role of users. There were 14 topics that were chosen for performing the evaluation and some of these were animals, celebrities, cinema, current affairs etcetera. Each of the topics was assigned to a different student. The search queries were chosen by students in French language. The inputs were then entered into different search engines on the same day. The web search engine settings were modified to return results in French.

There were a total of 4200 URLs that were gathered out of which 3450 URLs were unique. Each student was provided with a document which had a search query along with the corresponding URL. They were asked to evaluate each URL without being aware of corresponding search engine.

**Evaluation Criteria**

The Evaluation criteria on which URLs were graded by the students are as follows :-

- **Dead link**: This category of URL was evaluated and graded as zero if the website did not respond otherwise it was graded as one.
- **Pornographic link**: This characteristic was either graded zero or one depending upon whether a particular URL redirects the user to a pornographic website.
- **Topic**: This aspect of the URL was graded one if the result found was on the topic else it was graded zero.
- **Commercial site**: It was graded either zero or one depending on whether the Website being accessed is an e-commerce site or not.
- **Relevance**: It was graded on scale of 0-5. A zero grade means that the search result URL is off the topic and not relevant to the user.

5. **RESULTS**

Following are the results from the various methods specified above.

5.1 Results from Method One

i. 10% of the sampled pages were found to be dynamic in nature.

ii. Overlapping for static pages alone was found to be 10% higher than that for all pages.

iii. Weakening of URL matching criterion resulted in increased result overlap.

iv. Less than 1.4% of results overlapped for the four search engines under examination.

v. Each search engine seemed to index a fixed fraction of other search engines.

5.2 Results from Method Two

Some of the major results drawn by these authors are:

i. A majority of the results on the first result page for a specific query are unique to a particular search engine. This uniqueness is a consequence of the indexing and ranking techniques used by the search engines.

ii. The percentage of result overlap between the search results of different search engines decreases with an increase in the number of search engines.

![Figure 3: Variation of search results overlaps v.r.t variation in search engines count](image)

The data depicted in figure-2 is hypothetical and has been used to show decrease in search results overlap with increase in the number of search engines.

*Our suggestion*- This result can be used to understand that the coverage for different search engines varies in several aspects such as coverage size, languages covered number of sponsored and non-sponsored links etcetera.

iii. The search result provided by a meta-search engine like Dogpile.com is a collection of first page unique results of different single-source search engines.

5.3 Results from Method Three

**Composition of Web Indexes**: The web indexes created by the search engines depend upon the extent to which it covers the web, frequency with which it covers the database and also on the location of the web pages that are indexed. For instance, *AltaVista* indexed 2.5 million web pages while *Excite* indexed 1.5 million web pages and *Lycos* indexed 95% of the web resources.

*Search Capability*: This is one of the significant criteria responsible for determining a powerful search engine. This capability can be related to *Boolean logic, phrase searching, truncation, title searching* etc.
Table 1: Search capabilities of three search engines

<table>
<thead>
<tr>
<th>Search capabilities</th>
<th>AltaVista</th>
<th>Lycos</th>
<th>Excite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean search</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Proximity search</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Truncation search</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Field search</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case sensitivity</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concept search</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

As shown in Table 1 above, it can be observed that all the search engines provide the capability of Boolean searching. *Alta Vista* provides the option of Proximity search using NEAR operator as well as field search. *Excite* facilitates users to search a concept whereas only Lycos truncates the user’s input.

*Retrieval Performance:* It relates to precision and response time of a web search engine with respect to which it outputs the search results.

![Graph](image)

*Figure 4: Precision values versus Web search engines*

*AltaVista* not only showed the most precise results but also its response time to be less than 3 seconds. While for other search engines such as *Excite*, response time was found to be between 3 to 5 seconds.

*Output option:* This evaluation validates search engines for the quality and quantity of the output options offered to the user. A web search engine may offer thousands of results but they may not be relevant in accordance with the user’s needs.

*User Effort:* A user friendly interface provides clarity of interaction with search engines. Well prepared user guides in the form of documents add to the understanding and ease of the user. Comparative results of the search engines are shown in the graph below on a scale of 1-5.

5.4 Results from Method Four

*Hit Count Estimates:* The observations reflected that Hit Count Estimates (HCEs) of *Google*, *Yahoo* and *Live Search* closely correlate with each other. It was also observed that *Yahoo* search results showed less degree of correlation with *Google* and *Live Search* results, because of the fact that *Yahoo* automatically corrected some user errors. Thus, *Yahoo* displayed results after performing search on both corrected as well as uncorrected user queries. Following table shows the degree of correlation.

![Table 2: Correlation between Search engines](image)

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Yahoo!</th>
<th>Live Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>-</td>
<td>0.80</td>
<td>0.96</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>0.80</td>
<td>-</td>
<td>0.83</td>
</tr>
<tr>
<td>Live Search</td>
<td>0.96</td>
<td>0.83</td>
<td>-</td>
</tr>
</tbody>
</table>

*Number of URLs returned:* This characteristic showed the number of URLs that were returned for a particular search query. The search results so obtained showed that *Yahoo* and *Google* returned maximum number of URLs at the time when *Live Search* returned comparatively less number of URLs.
of dead links were directed to www.amazon.fr. It was also observed that maximum numbers of dead links. It was also observed that maximum numbers of dead links was studied for above six search engines and it was graded '1' if it did not respond to the user's request during the time indexing was performed. A URL was graded '0'.Search engines were analysed to any temporary problem or in case the page disappears 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. The proportion of dead links was studied for above six search engines and it was Exalead that offered the highest number of dead links. It was also observed that maximum numbers of dead links were directed to www.amazon.fr.

5.5 Results from Method Five

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. The proportion of dead links was studied for above six search engines and it was Exalead that offered the highest number of dead links. It was also observed that maximum numbers of dead links were directed to www.amazon.fr.

Pornographic link: There can be instances when the URL for search results redirects the user to a pornographic website. Due to availability of Safe Search option it was found that only two search engines (Voila and MSN) redirected the user to pornographic websites.

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 Voila 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 PriceMinister were the most commonly returned commercial websites. MSN was observed to be the only search engine that returned least number of links to commercial sites (7.1 %). Google and Yahoo were closely associated with Amazon while Voila with eBay and PriceMinister.

![Figure 6: Correlation for the Number of URLs returned](image1)

X-> Degree of correlation
Y-> Web search engines

Above representation shows that Yahoo and Live Search showed a high degree of correlation (0.84) while Google and Live Search showed a less degree of correlation (0.68) when measuring the number of URLs returned.

Number of domains returned: It relates to the domain of the URLs obtained for each search query. For example, in the URL www.java2s.com/tutorial , domain name is java2s.com.Yahoo emerged out be clear winner in this feature of returning the maximum number of domains in its results URLs. The results showed highest degree of correlation among Yahoo and Live Search (0.83).

Number of TLDs returned: TLD refers to the top-level domains mined from the list of URLs returned as search results. This feature tests the extent to which the search results are internationally dispersed. Yahoo Search results showed 10% more top level domains as compared to others.

![Figure 7: Proportion of off topic for various search engines](image2)

Relevance: It shows the degree to which the search results are related to the user’s needs. It was found that no search engine achieved a pass grade of 2.5; however Google and Yahoo produced most relevant results.

The relevance of search results varied with the position of the URL on web pages. Almost all web search engines showed a decreasing trend in the relevancy of the search result as we moved from top to bottom, except Voila which had least relevant results in its topmost position.

6. SUMMARY & CONCLUSIONS

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 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. Search engines with a single-source have low web coverage in comparison to a meta-search engine. Therefore, meta-search engine such as Dogpile.com is a better alternative to the use of multiple single-source search engines for the same query as it helps user to see all the relevant results from multiple web databases by feeding the search query only once. Yahoo offers a huge number of search results as it searches for both keywords with typo errors and rectified search
keywords. Documentation and user friendly interface reduces user’s efforts to perform a quest. Lycos was found to be good in terms of documentation and usability of interface. 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 single-source 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 relevancy to the search query except Voila which follows the reverse order. As per the Jean Veronice’s study, Google and Yahoo had a high user satisfaction index. Yahoo displayed least number of irrelevant topics. The above 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.

7. FUTURE WORK
Future work can be performed in extension of the current work to cover the limitations of the current work such as comparing search engines with sample data in more than 1 language, considering dynamic pages, including more number of meta-search engines, considering first two result pages for data collection instead of one, performing multiple types of URL matches to estimate search result overlap.

8. ACKNOWLEDGEMENT
We owe our thanks to Jan Gustafsson and Gordana Dodig-Crnkovic for their valuable inputs and guidance to present our work. An overview of the comparative studies for search engines performed by various researchers has been presented in this paper. Our contribution is limited to the advices enlisted and a comprehensive presentation of the work performed by other researchers to help users identify a search engine that meets their needs to a good extent.

9. REFERENCES


