IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 04, Issue 02 (February. 2014), ||V1|| PP 01-12

Performance Evaluation of Selected Search Engines

[1]Ajayi, Olusola Olajide, [2]Elegbeleye, Damilola Matthew
Department of Computer Science, Adekunle Ajasin University Akungba-Akoko, Ondo State, Nigeria
Department of Computer Science, Adekunle Ajasin University Akungba-Akoko, Ondo State, Nigeria

Abstract: - Search Engines have become an integral part of daily internet usage. The search engine is the first stop for web users when they are looking for a product. Information retrieval may be viewed as a problem of classifying items into one of two classes corresponding to interesting and uninteresting items respectively. A natural performance metric in this context is classification accuracy, defined as the fraction of the system's interesting/uninteresting predictions that agree with the user's assessments. On the other hand, the field of information retrieval has two classical performance evaluation metrics: precision, the fraction of the items retrieved by the system that are interesting to the user, and recall, the fraction of the items of interest to the user that are retrieved by the system. Measuring the information retrieval effectiveness of World Wide Web search engines is costly because of human relevance judgments involved. However, both for business enterprises and people it is important to know the most effective Web search engines, since such search engines help their users find higher number of relevant Web pages with less effort. Furthermore, this information can be used for several practical purposes. This study evaluates the performance of three Web search engines. A set of measurements is proposed for evaluating Web search engine performance.

I. INTRODUCTION

Search engines play a pivotal role in the process of retrieving information from the Web. A Web search engine is an information retrieval system (Salton & McGill, 1983), which is used to locate the Web pages relevant to user queries. A Web search engine contains indexing, storage, query processing, spider (or crawler, robot), and user interface subsystems.

The indexing subsystem aims to capture the information content of Web pages by using their words. During indexing, frequent words (that, the, this, etc.), known as stop words, may be eliminated since such words usually have no information value. Various statistics about words (e.g., number of occurrences in the individual pages or in all of the indexed Web pages) are usually stored in an inverted file structure. The storage technique deals with how to store the index data, that is whether information should be data compressed or filtered, the spider subsystem brings the pages to be indexed to the system. However, for Web users a search engine is nothing but its user interface that accepts queries and presents the search results.

When the user gives a Query, as a response, a Search engine returns a list of relevant results ranked in order. As a human, it is the tendency of the user to use top-down approach of the list displayed by the Search Engine and examines one result at a time, until the required information is found. The results on different subtopics or meanings of a query will be mixed together in the list, thus implying that the user may have to sift through a large number of irrelevant items to locate those of interest. On the other hand, there is no way to exactly figure out what is relevant to the user given that the queries are usually very short and their interpretation is inherently ambiguous in the absence of a context.

The growth of the World Wide Web is a unique trend. Four years after the Webs birth in 1990, a million or more copies of the first well-known Web browser, this growth was a result of the exponential increase of Web servers and the number of Web pages made accessible by these servers. In 1999 the number of Web servers was estimated at about 3 million and the number of Web pages at about 800 million (Lawrence & Giles, 1999).

There are millions of Web users and about 85% of them use search engines to locate information on the Web (Kobayashi & Takeda, 2000). It is determined that search engine use is the second most popular Internet activity next to e-mail (Jansen & Pooch, 2001). Due to high demand there are hundreds of general purpose and thousands of specialized search engines (Kobayashi & Takeda, 2000; Lawrence & Giles, 1999). Examples of General purpose search engines are, Google, Netscape, Lycos, AlltheWeb, AltaVista, HotBot, InfoSeek, Lycos, MSN, Netscape, and Yahoo, while that of specialized search engines are Adam, Labyrinth, Voice of the shuttle, Eric, FinAid, Medseek, Go Ask Alice, FindLaw, LawGure etc.

The motivation of this study stem from the fact that Evaluation of search engines may need to be done often due to changing needs of users or the dynamic nature of search engines (e.g., their changing Web coverage and ranking technology) and therefore it needs to be efficient. More also, this research was prompted by the large amount of search traffic directed to a handful of Web search engines, even though many have similar interfaces and performance, there could be many possible avenues to investigate. This study is majorly concerned with comparative analysis of the performance of search engines, the text retrieval performance in static document collections.

This study provides statistically, significant consistent results compared to human-based evaluations (Interview and questionnaire). The effectiveness of search engines measured in terms of precision and recall. The Methodology further emphasizes the elicitation of genuine information needs from genuine users, as well as relevant judgments made by those same individuals.

This study proposes a comprehensive survey on four selected Search Engines and their performance in the process of information retrieval from the Web, and not all search engines. Since the scope of Internet search engines, such as Google, yahoo, ask and bing is the entire Web, ideally, the experiments would be conducted using all the data on the Web is obviously impossible. What is needed is a relatively small subset that resembles the Web as a whole.

The scope of the study it to established a set of standard measurement for the evaluation of information retrieval system aside from the most commonly used recall and precision include performance stability in any Web search engine evaluation.

Criteria for Evaluating Search Engine's Performance

Many publications compare or evaluate Web search engines (e.g. Notess, 2000). Perhaps the best known of these are Search Engine Watch (http://www.searchenginewatch.com) and Search Engine Showdown (http://www.searchengineshowdown.com). However, many of these publications did not employ formal evaluation procedures with rigorous methodology. Some papers that describe advances in search algorithms gave anecdotal evidence instead of formal evaluations (e.g. Brin & Page, 1998). Decades of research, from the classic Cranfield experiments to the ongoing TREC, have established a set of standard measurements for the evaluation of information retrieval systems. Only studies that used formal evaluation procedures are reviewed below.

Recall in Search Engine Performance Evaluation

While recall and precision have been the standard evaluation criteria for information retrieval since the Cranfield tests, recall has always been a difficult measure to calculate because it requires the knowledge of the total number of relevant items in the collection. This was possible in small laboratory studies such as the Cranfield tests. It becomes increasingly difficult as collection size grows. This problem is more acute in the Web environment. Chu and Rosenthal's Web search engine study omitted recall as an evaluation measure because they consider it ‘‘impossible to assume how many relevant items there are for a particular query in the huge and ever changing Web systems’’ (Chu & Rosenthal, 1996, p. 127). Gwizdka and Chignell (1999) acknowledged the difficulty of calculating recall on the Web for the same reason. They did not include recall in their recommended measures of search engine evaluation. TREC (Text Retrieval Conference) Web track used a pooling method to find ‘‘all’’ relevant documents to calculate recall. It was assumed that documents not in the pool were not relevant (Voorhees & Harman, 2001, p. 4). The study reported in this paper proposes a modified measurement similar to that of TREC in principle, i.e. it uses pooling, but it employs a continuous relevance ranking (from the most relevant to the least relevant) rather than the binary relevance judgment used in TREC. It should be noted that TREC's use of pooling method to calculate recall has been criticized because recall calculated this way ‘‘will be higher-perhaps substantially higher-than what they actually are’’ (Blair, 2002, p. 449). Organizers of TREC agrees that the recall calculated this way would be higher but argues that this ‘‘relative recall’’ is valid for comparing the relative performance of different systems (Saracevic, Voorhees, & Harman, 2003). The same argument applies to this study where the comparison of relative performance of search engines is the purpose.

Precision in Search Engine Performance Evaluation

Precision is always reported in formal information retrieval experiments. However, there are variations in the way it is calculated depending on how relevance judgments were made. ‘‘TREC almost always uses binary relevance judgment-either a document is relevant to the topic or it is not’’ (Voorhees & Harman, 2001, p. 4). Hawking, Craswell, Bailey, and Griffiths (2001), Hawking, Craswell, Thistlewaite, and Harman (1999) used binary relevance judgment and calculated traditional recall and precision measures at various cut-off levels. Recognizing the fact that there could be a degree of relevance, many studies used multi-level rather than binary relevance judgments. For example, Chu and Rosenthal (1996) used a three-level relevance score (relevant, somewhat relevant, and irrelevant) while Gordon and Pathak (1999) used a four-level relevance judgement (highly relevant, somewhat relevant, somewhat irrelevant, and highly irrelevant). Both studies calculated the traditional recall and precision to compare search engines. Ding and Marchionini (1996) employed a six-point scale and took into consideration links to other relevant documents. They calculated three different types of precision using the six-point scale. One of the most important features of Web search engines is result ranking, without which it is simply impossible for a user to sift through the hundreds or even tens of thousands of items retrieved. ‘‘Results ranking has a major impact on users_ satisfaction with Web search engines and their success in retrieving relevant documents. Yet, little research has been done in this area’’.

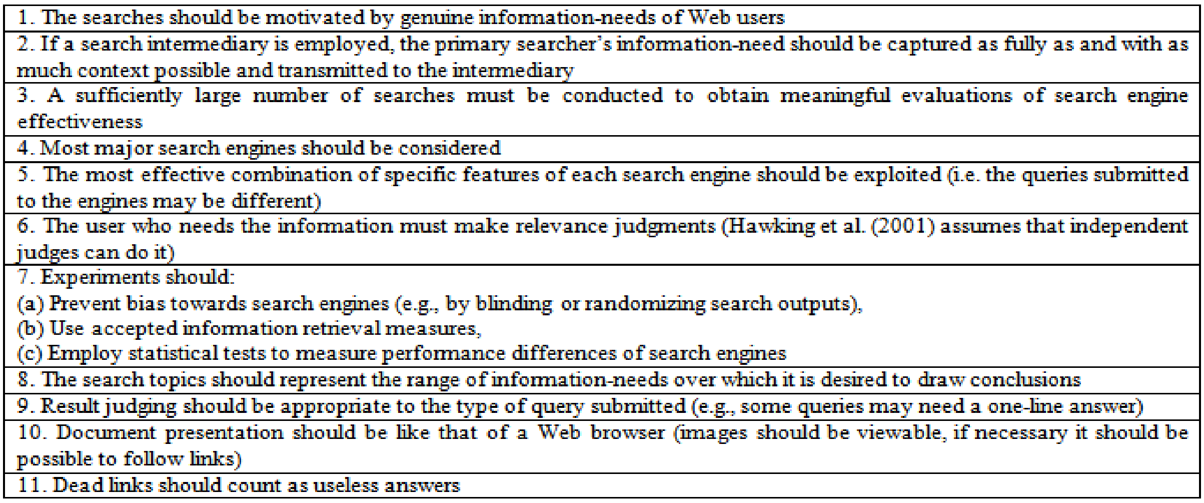
(Courtois & Berry, 1999). Gwizdka and Chignell (1999) acknowledged the importance of result ranking in Web information retrieval and developed the ‘‘differential precision’’ to measure the quality of ranking produced by search engines. However, their approach is still based on a four point scale relevance judgment. Similarly, Su, Chen, and Dong (1998) used a five-point relevancy scale and then correlated these rankings with the rankings returned by search engines. The problem of using these discrete relevance scores is that two or more documents can easily receive the same score and be tied in the ranking. This scoring system is therefore not very effective in evaluating ranking results where there are usually no ties. The study reported here proposes a continuous ranking (from most relevant to least relevant) of the document set instead of the discrete relevance judgments. The correlation between human ranking and search engine ranking can be calculated instead of the traditional measure of precision.

Human Relevance Judgments

Not all search engine studies used human relevance judgment as the basis of evaluation, probably due to the difficulty and expense of such efforts. Courtois and Berry (1999) studied the first 100 items retrieved by five search engines. They did not use human relevance judgment, which would be extremely difficult to do given the total number of items for which relevance judgments need to be made in their study. Instead, they used a computer program to automatically check the location, proximity, etc. of the search terms in the retrieved documents and used this information to compare the search engines in the study.

When human relevance judgment was used, there was a variation in who makes the judgment. TREC leaves relevance judgments to experts or to a panel of experts (Voorhees & Harman, 2001). In other studies, relevance judgments were made by the researchers themselves (e.g. Chu & Rosenthal, 1996). Gordon and Pathak (1999) emphasized that relevance judgments can only be made by the individual with the original information need.

The Gordon and Pathak (1999) study measures the performance of eight search engines using 33 information-needs. For measuring performance it calculates recall and precision at various document cut-off values (DCVs) and uses them for statistical comparisons. Intermediaries prepare the queries from the information-need descriptions of real users. The query preparation is done iteratively to achieve the best performance of individual search engines and therefore each individual search engine query used for the same information-need may be different. The user who originated the search did the assessment of the top 20 results of each search engine. The findings of the study indicate that absolute retrieval effectiveness is low and there are statistical differences in the retrieval effectiveness of search engines. The study recommends seven features to maximize the accuracy and informative content of such studies (see Table 2). The study reported in Hawking et al. (2001) evaluates the effectiveness of 20 search engines using TREC-inspired methods with 54 queries taken from real Web search logs. The performance measures used include precision at various DCVs, mean reciprocal rank of first relevant document, and TREC-style average precision. Recall has not been used. Statistical testing reveals high inter-correlations between performance measures and significant differences between performances of search engines. Despite the time difference among the results of this study and those of Gordon and Pathak (1999) a high level of correlation is observed. This study proposes some more features F. Can et al. / Information Processing and Management 40 (2004) 495–514 497



1. The searches should be motivated by genuine information-needs of Web users
2. If a search intermediary is employed, the primary searcher’s information-need should be captured as fully as and with as much context possible and transmitted to the intermediary
3. A sufficiently large number of searches must be conducted to obtain meaningful evaluations of search engine effectiveness
4. Most major search engines should be considered
5. The most effective combination of specific features of each search engine should be exploited (i.e. the queries submitted to the engines may be different)
6. The user who needs the information must make relevance judgments (Hawking et al. (2001) assumes that independent judges can do it)
7. Experiments should: (a) Prevent bias towards search engines (e.g., by blinding or randomizing search outputs), (b) Use accepted information retrieval measures, (c) Employ statistical tests to measure performance differences of search engines
8. The search topics should represent the range of information-needs over which it is desired to draw conclusions
9. Result judging should be appropriate to the type of query submitted (e.g., some queries may need a one-line answer)
10. Document presentation should be like that of a Web browser (images should be viewable, if necessary it should be possible to follow links)
11. Dead links should count as useless answers

Table 1 Desirable features of Web search evaluation according to Gordon and Pathak (1999): features 1–7 and Hawking et al. (2001): features 8–11

II. DESIGN SPECIFICATION

The Underlying Logics

A typical search engine is composed of three pipelined components (Arasu, Cho,Garcia- Molina, & Raghavan, 2001): a crawler, an indexer, and a query processor. The crawler component is responsible for locating, fetching, and storing the content residing within the Web. The downloaded content is concurrently parsed by an indexer and transformed into an inverted index (Tomasic, Garcia-Molina, & Shoens, 1994; Zobel, Moffat, & Sacks-Davis, 2002), which represents the downloaded collection in a compact and efficiently queryable form. The query processor is responsible for evaluating user queries and returning to the users the pages relevant to their query. The web crawling (downloading of web pages) is done by several distributed crawlers. There is an URL server that sends lists of URLs to be fetched to the crawlers. The web pages that are fetched are then sent to the store server. The store server then compresses and stores the web pages into a repository. Every web page has an associated ID number called a docID which is assigned whenever a new URL is parsed out of a web page. The indexing function is performed by the indexer and the sorter. The indexer performs a number of functions. It reads the repository, uncompresses the documents, and parses them. Each document is converted into a set of word occurrences called hits. The hits record the word, position in document, an approximation of font size, and capitalization. The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index. The indexer performs another important function. It passes out all the links in every web page and stores important information about them in an anchors file. This file contains enough information to determine where each link points from and to, and the text of the link. The URLresolver reads the anchors file and converts relative URLs into absolute URLs and in turn into docIDs. It puts the anchor text into the forward index, associated with the docID that the anchor points to. It also generates a database of links which are pairs of docIDs. The links database is used to compute PageRanks for all the documents.

On a less complicated level, search engines could simply be described as suites of computer programs interacting and communicating with each other. Different, or various terms for the particular components are used by search engines in their development and research. The components are: crawler/spider module, repository/database module, indexer/link analysis module, retrieval/ranking module, user query interface, crawler/spider module. Other components include: document routing, filtering, and selective dissemination systems, text clustering and categorization systems, summarization systems, information extraction systems, topic detection and tracking systems, expert search systems, question answering systems, multimedia information retrieval systems.

Algorithms For Search Engine

Unique to every search engine, and just as important as keywords, search engine algorithms are the why and the how of search engine rankings. Basically, a **search engine algorithm** is a set of rules, or a unique formula, that the search engine uses to determine the significance of a web page, and each search engine has its own set of rules. These rules determine whether a web page is real or just spam, whether it has any significant data that people would be interested in, and many other features to rank and list results for every search query that is begun, to make an organized and informational search engine results page. The algorithms, as they are different for each search engine, are also closely guarded secrets.

Each search engine has its own way of doing things. Discuss below is a typical algorithm for designing a search engine.

1. User interface send query to the search engine
2. Interface receive the query and parser parses the query
3. Convert the words into words IDs.
4. Seek to start searching through the doc-list in the short barrel for every word
5. Scan through the doc-list until there is a document that matches all the search terms
6. Compute the rank of that document for the query.
7. If we are in the short barrels and at the end, seek to start searching of the doc-list in the full barrel for every word and go to step 4
8. If we are not at the end of any doc-list go to step 4. Sort the documents that have matched by the rank and return to the top

Mathematical Algorithm for Designing Search Engine

The following design standards can be use, as design guidelines specified for search engines.

S1: Programming Standard

- Architecture: Information related to hardware abstraction, software requirements and user interface design

- Implementation: Actual system specifications, operating system and other software selection criterion, and application of software engineering procedures to realize the Search Engine as a product.

S2: Design Standard

- Recency: Ensuring that each document d_i at SE has same timestamp as the original web site.
- Relevancy: Exact document(s) d_i to be retrieved by SE that the user is looking for.

S3: Data Acquisition Standard

Robots/Spiders: Design, implementation and deployment mechanism and principles.

S4: Data Submission Standard

- Submission Types: User submission guidelines for directory search as well as for paid services.

S5: Data Processing Standard

- Page Ranking: Document the existing page rank techniques and look for improvisation.
- Indexing Techniques: Research and documentation of indexing techniques.

S6: Information Retrieval Standard

- Search Algorithms: Detail the data structures, analyze the complexity of various algorithms and suggest the suitability for various domains.
- Scope for Intelligence: Identify the scope for the intelligence and incorporate appropriately.

S7: User Interface Standard

- User Interface: Study and list the applicability of HCI findings for elegant, intuitive interface.
- Natural Language Processing: A natural requirement for human interface.
- Multimedia Search: Enlist various standards for multimedia and assess the ability of SEs to incorporate them.

The benchmark metric is a weighted average of the following parameters. We give two level weights for the parameters.

Benchmark Metric (BM) for SE = W1 * UP + W2 * SP + W3 * QP,

Where,

UP = w1*UPP1+ w2*UPP2+ w3*UPP3+ w4*UPP4+ w5*UPP5,

SP = w1*SPP1+ w2*SPP2+ w3*SPP3+ w4*SPP4 and

QP = w1*QPP1+ w2*QPP2+ w3*QPP3

And the details of the parameters are given below.

User Perspective (UP):

- UPP1: Completeness: Indexing the entire available web repository.
- UPP2: Precision: The ratio of relevant documents to retrieved documents
- UPP3: Recall: The ratio of retrieved documents to relevant documents
- UPP4: Recency: Maintaining the same timestamp at SE and the actual website.
- UPP5: Relevancy: Retrieving the exact document(s) the user is looking for.

System Perspective (SP):

Hardware:

- SPP1: Infrastructure: Hardware and Software requirements and cost optimization.

Software:

- SPP2: Indexing Algorithms: Efficiency, space and time complexity.
- SPP3: Ranking Algorithms: Rank technology, efficiency and effectiveness.
- SPP4: Robots Specification: Design, implementation, overhead and download capacity.

Query Perspective (QP):

- QPP1: Natural Language Processing (NLP): A natural interface to accept queries, for the users to specify the queries.(.pdf. doc. ppt. etc.)
- QPP2: Multimedia Content Analysis: Ability to accept the query in multimedia format, search and the efficiency of the multimedia content classification an search algorithms are evaluated.
- QPP3: Adjacency, Synonyms and Stemming: Ability to search for synonyms to yield relevant results is quantified by this parameter.

III. IMPLEMENTATION

Measuring Performance of Search engines

Retrieval measures are used to measure the performance of IR systems and to compare them to one another. The main goal for search engine evaluation is to develop individual measures (or a set of measures) that are useful for describing the quality of search engines. Performance and quality of information retrieval are important components of search engine evaluation model.

The performance and retrieval quality attributes can be measured using a specialized benchmarking tool. In this case the drive workload must reflect the interest of general public, and we used the statistics of most frequent requests obtained from traffic monitors of major search engines. For example is the workload for a single query “information systems” search on three search engine on 29/10/2012

Workload	Retrieved URL	Query Type
Google	762,000,000	Double Word
Yahoo	325,000,000	Double Word
Bing	21,100,000	Double Word

Table 2: Workload for query

Workload has strong semantic component and the quality of information retrieval can be fully analyzed only by experts in a specific area. For example, the recall of a query about “andness” should be evaluated by decision analysts, and the recall of a query about “Rituximab” should be evaluated by medical experts. A complete analysis of performance and quality of search engines for workload cannot be done automatically using a benchmark tool. Therefore, an automatic analysis of performance and quality of search is reasonable for workload in our tool for search engine benchmarking and for measurement of quality of information retrieval (IR). Following are main design goals:

1. Measurement of Precision
2. Measurement of search engine coverage
3. Search engine stability measurement

Precision measures the ability of Search Engine to produce only relevant results. Precision is the ratio between the number of relevant documents retrieved by the system and the total number of documents retrieved. An ideal system would produce a precision score of 1, i.e. every document retrieved by the system is judged relevant. Precision is relatively easy to calculate, which mainly accounts for its popularity. But a problem with precision in the search engine context is the number of results usually given back in response to typical queries. In many cases, search engines return thousands of results. In an evaluation scenario, it is not feasible to judge so many results. Therefore, cut-off rates (e.g. 20 for the first 20 hits) are used in retrieval tests.

The coverage of a search engine can be determined as the total number of pages returned by the search engine.

Stability measurements are to examine the stability of search engine performance over time. The measurements are:

(1) The stability of the number of pages retrieved;

(2) the number of pages among the top 20 retrieved pages that remain the same in two consecutive tests over a short time period (e.g. a week apart); and

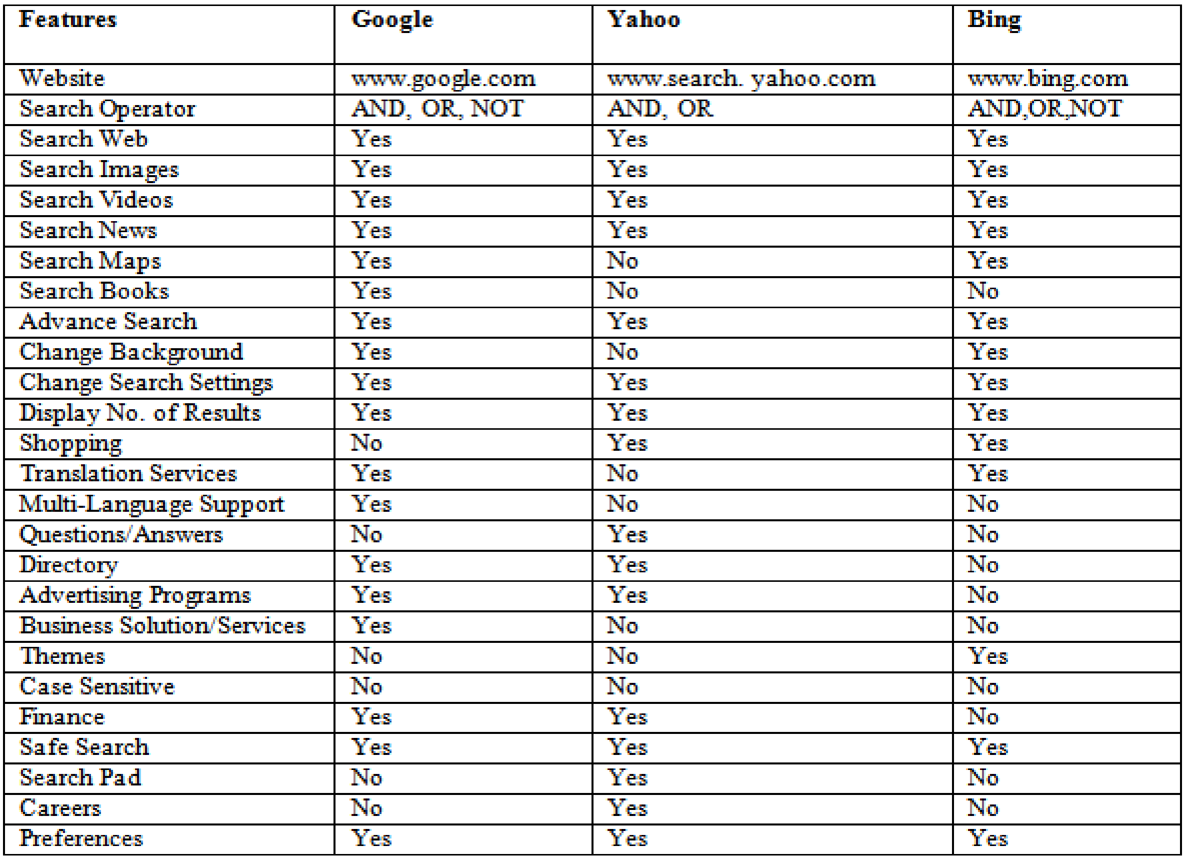
(3) The number of pages among the top 20 retrieved pages that remain in the same ranking order in two consecutive tests over a short time period (e.g. a week apart). Essentially, a series of searches needs to be performed on a search engine over a period of time (e.g. one search a week over a 10-week period). The number of pages retrieved needs to be recorded. The top 10 pages retrieved is compared with those from the previous search to determine

(a) If the pages retrieved are the same;

(b) If the ranking of the pages is the same.

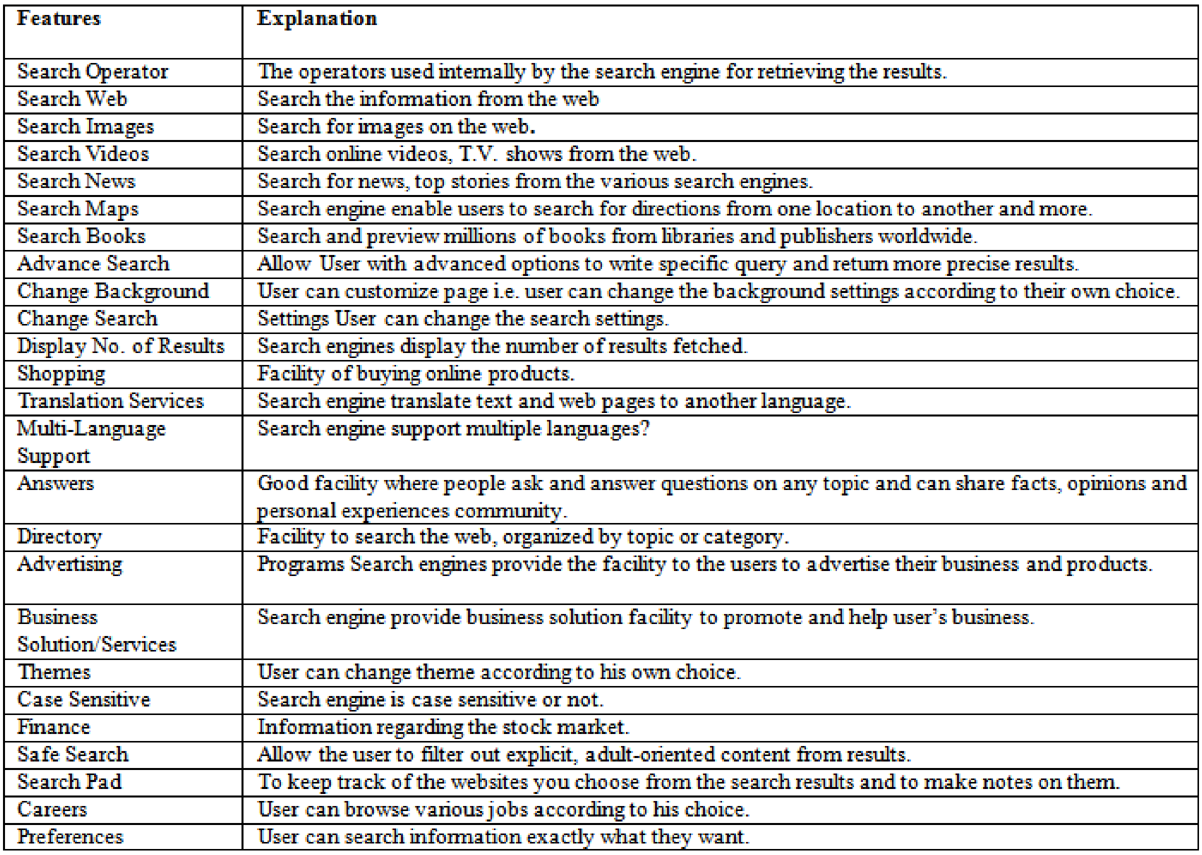
The examination of the top 10 pages is carried out because it will be extremely time consuming to compare each page in the entire retrieved set, which is typically very large.

For this study, three search engines were chosen. These are Google, Yahoo, and Bing. The criterion was the popularity of the search services. According to Search Engine Watch, a search industry news site, the major international search engines are Google, Yahoo, MSN. (Sullivan, 2007). This research compares these search engines on the basis of the features provided by them on the home page as well as after displaying the results. Some of the basic features of these search engines are indicated in Table 3 and explanation of each feature in Table 4.



Features	Google	Yahoo	Bing
Website	www.google.com	www.search. yahoo.com	www.bing.com
Search Operator	AND, OR, NOT	AND, OR	AND,OR,NOT
Search Web	Yes	Yes	Yes
Search Images	Yes	Yes	Yes
Search Videos	Yes	Yes	Yes
Search News	Yes	Yes	Yes
Search Maps	Yes	No	Yes
Search Books	Yes	No	No
Advance Search	Yes	Yes	Yes
Change Background	Yes	No	Yes
Change Search Settings	Yes	Yes	Yes
Display No. of Results	Yes	Yes	Yes
Shopping	No	Yes	Yes
Translation Services	Yes	No	Yes
Multi-Language Support	Yes	No	No
Questions/Answers	No	Yes	No
Directory	Yes	Yes	No
Advertising Programs	Yes	Yes	No
Business Solution/Services	Yes	No	No
Themes	No	No	Yes
Case Sensitive	No	No	No
Finance	Yes	Yes	No
Safe Search	Yes	Yes	Yes
Search Pad	No	Yes	No
Careers	No	Yes	No
Preferences	Yes	Yes	Yes

Table 3: Features available in the Three Selected Search Engines



Features	Explanation
Search Operator	The operators used internally by the search engine for retrieving the results.
Search Web	Search the information from the web
Search Images	Search for images on the web.
Search Videos	Search online videos, T.V. shows from the web.
Search News	Search for news, top stories from the various search engines.
Search Maps	Search engine enable users to search for directions from one location to another and more.
Search Books	Search and preview millions of books from libraries and publishers worldwide.
Advance Search	Allow User with advanced options to write specific query and return more precise results.
Change Background	User can customize page i.e. user can change the background settings according to their own choice.
Change Search	Settings User can change the search settings.
Display No. of Results	Search engines display the number of results fetched.
Shopping	Facility of buying online products.
Translation Services	Search engine translate text and web pages to another language.
Multi-Language Support	Search engine support multiple languages?
Answers	Good facility where people ask and answer questions on any topic and can share facts, opinions and personal experiences community.
Directory	Facility to search the web, organized by topic or category.
Advertising	Programs Search engines provide the facility to the users to advertise their business and products.
Business Solution/Services	Search engine provide business solution facility to promote and help user's business.
Themes	User can change theme according to his own choice.
Case Sensitive	Search engine is case sensitive or not.
Finance	Information regarding the stock market.
Safe Search	Allow the user to filter out explicit, adult-oriented content from results.
Search Pad	To keep track of the websites you choose from the search results and to make notes on them.
Careers	User can browse various jobs according to his choice.
Preferences	User can search information exactly what they want.

Table 4: Explanation of various features mentioned in Table 3

Queries in the study were designed to test various search features including single word search, double word search, Boolean search and phrase search. All queries were English language. The five search topics and their corresponding search queries were:

1. Development communication. (Double word)
2. Communications. (Single word)
3. Performance and Evaluation (Boolean search)
4. School of information Technology (phrase search)
5. No knowledge is lost (phrase search)

A set of Web pages on each topic was retrieved using the three search engines and the top 10 pages retrieved by each engine were merged to form the set of pages to be ranked for that particular topic. It should be noted that ''pages'' in this paper refers to the individual Web pages (or hits) retrieved, not the pages of search results (usually 10 hits are presented in a search results page). The decision to limit to the top 10 pages was based on the fact that human subjects may not be able to reliably rank (rather than making binary decision of relevant vs. non-relevant) more than 30 pages. Previous studies of Web search engines have used the pooling of top 10 (Schlichting & Nilsen, 1996) and top 7 pages (Hawking et al., 2001).

For each query, the first 10 results were analyzed from each search engine. A cut-off value had to be used because of the very large results sets produced by the search engines. From user studies, it was made clear that users only look at the first few results and ignore the rest of the result list. The first10 results per search engine was examined for each of the five query and the relevant and non- relevant was recorded in the table below and the precision of the three search engines was calculated.

Search queries	Google		Yahoo		Bing	
	Relevant	Irrelevant	Relevant	Irrelevant	Relevant	Irrelevant
Q#1	6	4	5	5	3	7
Q#2	6	4	4	6	2	8
Q#3	7	3	6	4	5	5
Q#4	9	1	3	7	2	8
Q#5	5	5	4	6	3	7

Search queries	Google Precision	Yahoo Precision	Bing Precision
Q#1	0.6	0.5	0.3
Q#2	0.6	0.4	0.2
Q#3	0.7	0.6	0.5
Q#4	0.9	0.3	0.2
Q#5	0.5	0.4	0.3

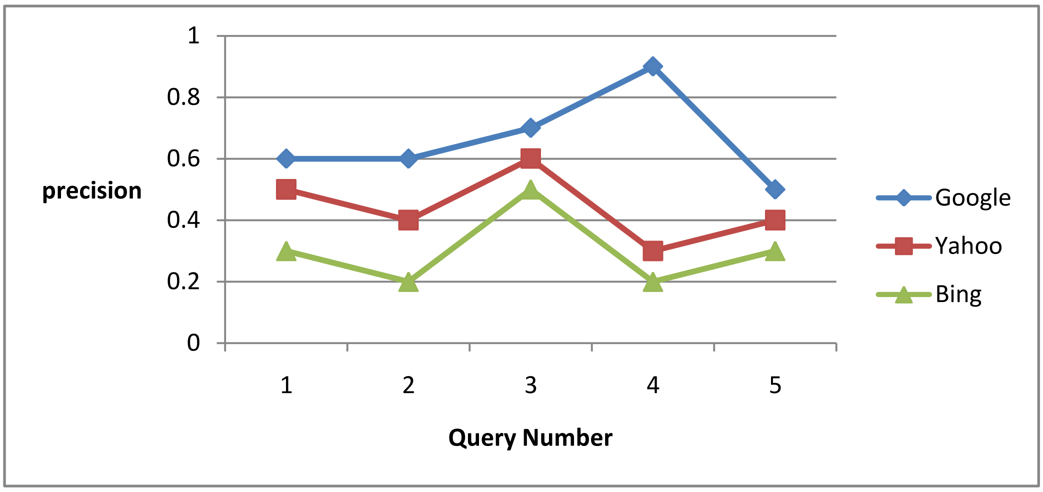


Fig. 1: Precision values for the three search engines

However, regarding only the top results i.e. only top ten pages, one can find clear difference between the engines. Google performs better than all other engines follow by yahoo and Bing comes up at the third position.

This may be one reason why the general public regards Google as the best search engine. When users only observe the first few results, they may easily come to that conclusion.

IV. MEASUREMENT OF COVERAGE

Search queries	Google Total Number of pages Returned	Yahoo Total Number of pages Returned	Bing Total Number of pages Returned
Q#1	462,000,000	1,280,000,000	575,000,000
Q#2	767,000,000	390,000,000	150,000,000
Q#3	803,000,000	959,000,000	17,800,000
Q#4	1,550,000,000	497,000,000	4,290,000,000
Q#5	188,000,000	9,320,000	5,490,000

The coverage measurement result proves that Google retrieved more result for each type of query at a single search.

Search Engine stability Measurement. The stability of search engine performance was first examined by looking at the stability of the number of pages retrieved over the 5-week period. The comparisons of the three engines in this regard are shown from Figs. 1–5 for the five queries respectively. The three Figures present the same pattern and thus same conclusions:

(1) Google consistently retrieved more pages than Yahoo and Bing; and

(2) Google remained very stable over the 5-week period without any sudden increase or decrease in the number of pages retrieved. In contrast, all the search engines increase the number of result page.

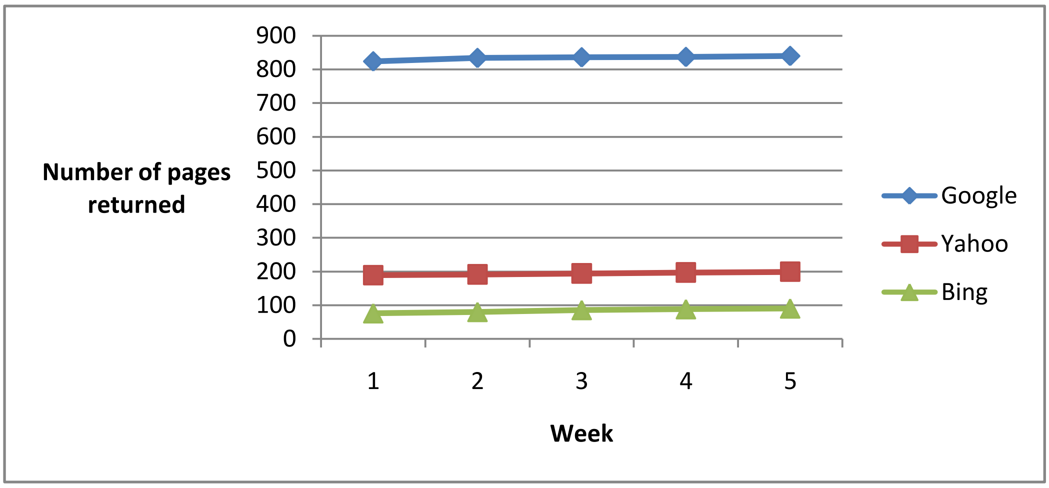


Fig. 2: Search results for query development communication

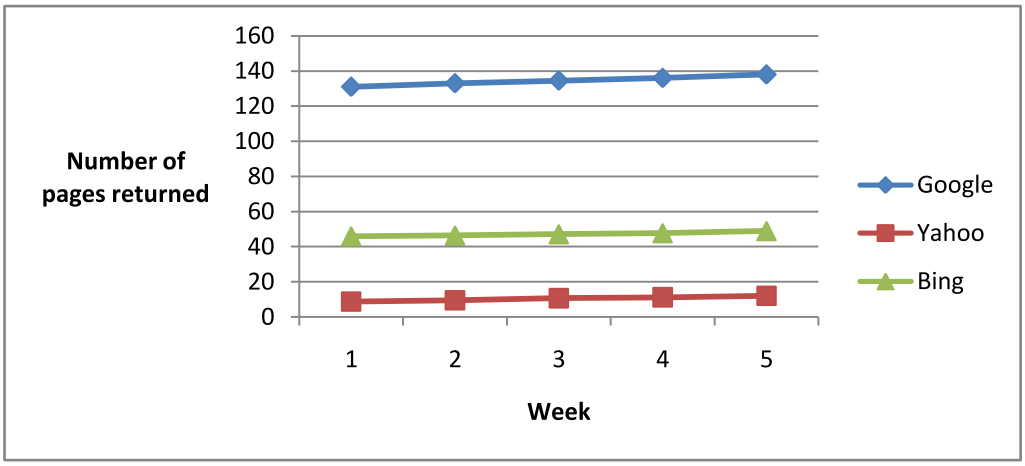


Fig. 3: Search results for query communication

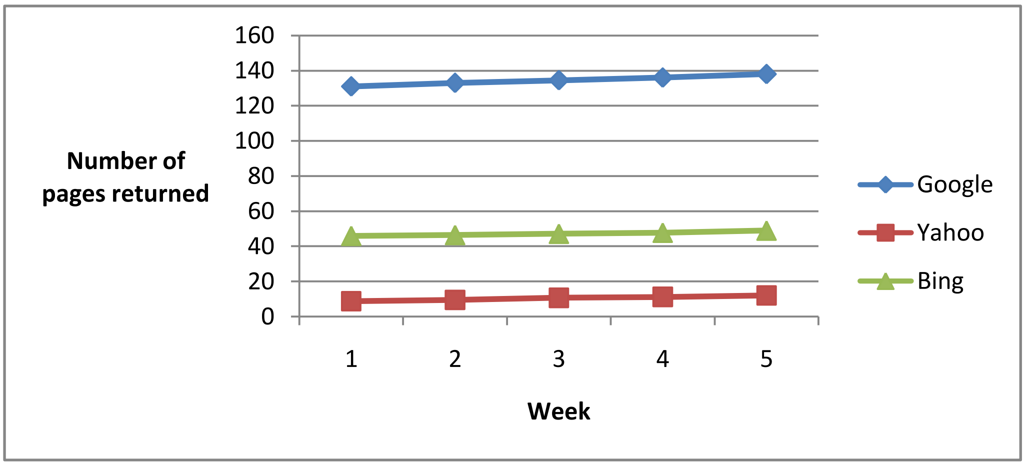


Fig. 4: Search Results for Query Performance and Evaluation

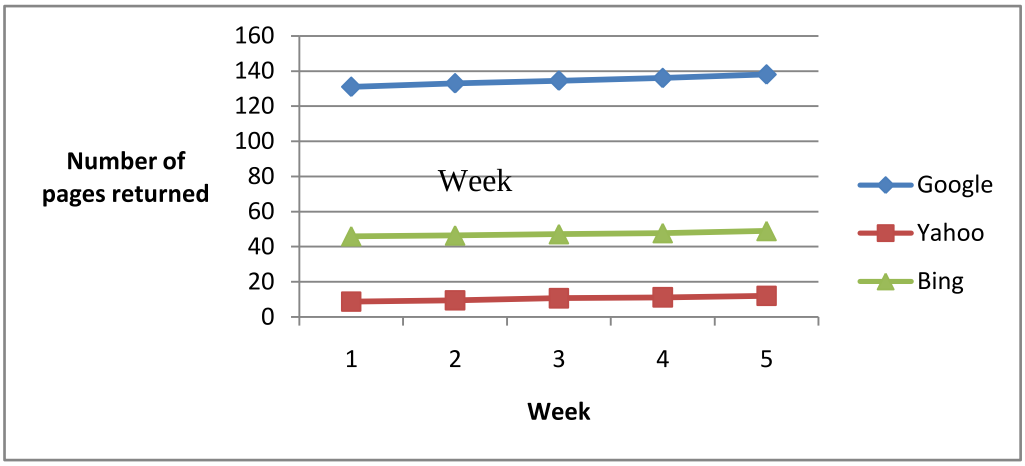


Fig. 5: Search Results for Query School of Information Technology

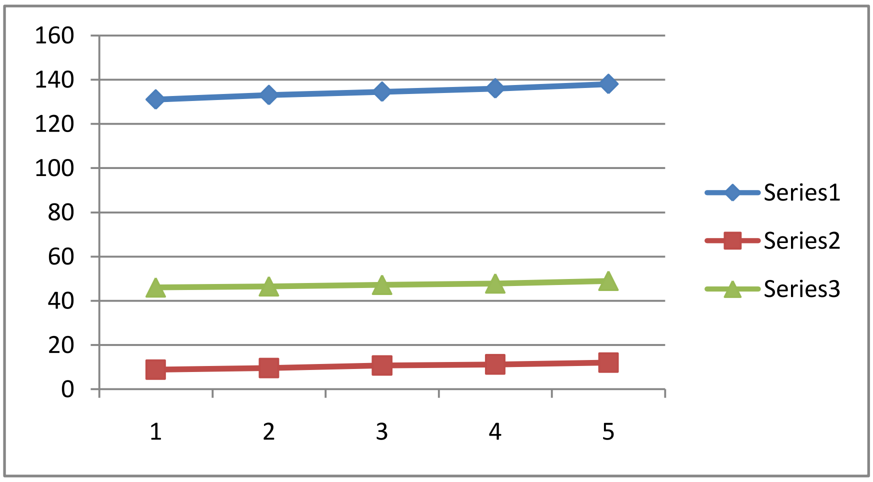


Fig. 6: Search Results for Query No knowledge is lost

V. CONCLUSION

Today, nobody knows the real performance or accuracy of search engines. There are several studies dealing with a single aspect of quality measurement, but none that tries to evaluate search engine quality as a whole. It must be noted that the purpose of the experiment in this study is to illustrate how the proposed measurements should be applied and to test the ability of the measurements in distinguishing search engine performance. The relative performance of the three engines upon the five test queries are only suggestive of their quality and should not be taken as conclusive evidence. To convincingly show that one engine is better than the other, a large and diverse set of queries is needed. Statistical tests should be performed on the large data set to determine if there are significant differences among the search engines examined.

One central question to our study was whether Google's perceived superiority in delivering relevant results could be confirmed by a systematic test. In this regard, we found that it makes no significant difference for a user to use one of the big search engines, Google or Yahoo.

When considering the first four results, Yahoo comes close to Google and then bing, and then Google leads in terms of precision for up to the seventh result. When considering more than seven results, the differences between Google and Yahoo are insignificant.

REFERENCES

[1] Bar-Ilan, J. (1999). Search engine results over time - a case study on search engine stability. Cyber-metrics, 2/3(1), Retrieved 15/12/2002 (http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html).

[2] Bar-Ilan, J. (2000). Evaluating the stability of the search tools Hot-Bot and Snap: A case study. Online Information Review, 24(6), 439–449.

[3] Bar-Ilan, J. (2002). Methods for measuring search engine performance over time. Journal of the American

[4] Society for Information Science and Technology, 53(4), 308–319.
Blair, D. C. (2002). Some thoughts on the reported results of TREC. Information Processing & Management, 38(3), 445– 451.

[5] Brin, S. & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7), 107–117.

[6] Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: a comparative study and evaluation methodology. In Proceedings of the 59th annual meeting of the American Society for Information Science (pp. 127– 135). Medford, NJ: Information Today.

[7] Courtois, M. P., & Berry, M. W. (1999). Results ranking in Web search engines. Online, 23(3), 39–46.

[8] Chowdhury, A. & Soboroff, I. (2002). Automatic evaluation of World Wide Web search services. In Proceedings of the ACM SIGIR conference, vol. 25 (pp. 421–422).

[9] Ding, W., & Marchionini, G. (1996). A comparative study of Web search service performance. In Proceedings of the 59th annual meeting of the American Society for Information Science (pp. 136–142). Medford, NJ: Information Today.

[10] Gordon, M. & Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. Information Processing and Management, 35(2), 141–180.

[11] Gwizdka, J., & Chignell, M. (1999). Towards information retrieval measures for evaluation of Web search engines. Retrieved 17/12/2002

[12] (http://www.imedia.mie.utoronto.ca/~jacekg/pubs/webIR_eval1_99.pdf).

[13] Hawking, D., Craswell, N., Thistlewaite, P., & Harman, D. (1999). Results and challenges in Web search evaluation. Computer Networks, 31(11–16), 1321–1330.

[14] Hawking, D., Craswel, N., Bailey, P., & Griffiths, K. (2001). Measuring search engine quality. Information Retrieval, 4(1), 33–59.

[15] Jansen, B. J., & Pooch, U. (2001). A review of Web searching and a framework for future research. Journal of the American Society for Information Science and Technology, 52(3), 235–246.

[16] Kobayashi, M., & Takeda, K. (2000). Information retrieval on the Web. ACM Computing Surveys, 32(2), 144–173.

[17] Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the Web. Nature, 400, 107–109.

[18] Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines – fluctuations in document accessibility. Journal of Documentation, 57(5), 623–651.

[19] Mowshowitz, A., & Kawaguchi, A. (2002). Assessing bias in search engines. Information Processing and Management, 35(2), 141–156.

[20] Notess, G. R. (2000). The never-ending quest: search engine relevance. Online, 24(3), 35–40.

[21] Notess, G. R. (2002). Search engine statistics: relative size showdown. Retrieved 02/09/2012 (http://

[22] www.searchengineshowdown.com/stats/size.shtml). Salton, G., & McGill, M. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

[23] Saracevic, T., Voorhees, E., & Harman, D. (2003). Letters to the editor. Information Processing & Management, 39(1), 153–156.

[24] Schlichting, C., & Nilsen, E. (1996). Signal detection analysis of WWW search engines. Presented at Microsoft's Designing for the Web: Empirical Studies Conference. Retrieved 17/12/2012

[25] (http://www.microsoft.com/usability/webconf/schlichting/schlichting.htm).

[26] Su, L. T., Chen, H., & Dong, X. (1998). Evaluation of Web-based search engines from the end-user's

[27] perspective: a pilot study. In Proceedings of the 61st annual meeting of the American Society for Information Science (pp. 348–361).

[28] Soboroff, I., Nicholas, C., & Cahan, P. (2001). Ranking retrieval systems without relevance judgments. In Proceedings of the ACM SIGIR conference, vol. 24 (pp. 66–73).

[29] Sullivan, D. (2002). Nielsen//Net Ratings: search engine ratings. Retrieved 23/12/2012 (http://www.searchenginewatch. com/reports/netratings.html).

[30] Sullivan, D. (2007), "Major search engines and directories", Search Engine Watch, available at http://searchenginewatch.com/showPage.html?page ¼ 2156221

[31] Thelwall, M. (2001). The responsiveness of search engine indexes. Cybermetrics, 5(1), Retrieved 17/12/2012 (http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html).

[32] Thelwall, M. (2002). In praise of Google: finding law journal Web sites. Online Information Review, 26(4), 271–272.

[33] Vaughan, L., & Hysen, K. (2002). Relationship Between Links to Journal Web Sites and Impact factors. Aslib Proceedings: New Information Perspectives, 54(6), 356– 361.

[34] Vaughan, L., & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to Journal Web sites? Journal of the American Society for Information Science and Technology, 54(1), 29–38.

[35] Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. Information Processing & Management, 40(4), 677–691.

[36] Voorhees, E. M., & Harman, D. (2001). Overview of TREC 2001. NIST Special Publication 500-250: The

[37] 10th text retrieval conference (TREC 2001) (pp. 1–15). Retrieved 17/12/2012

[38] (http://trec.nist.gov/pubs/trec10/papers/overview_10.pdf).