ELSEVIER

Information Processing and Management 40 (2004) 677–691

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

New measurements for search engine evaluation proposed and tested

Liwen Vaughan *

Faculty of Information and Media Studies, University of Western Ontario, London, Ont., Canada N6A 5B7

Received 22 January 2003; accepted 12 May 2003
Available online 19 June 2003

Abstract

A set of measurements is proposed for evaluating Web search engine performance. Some measurements are adapted from the concepts of recall and precision, which are commonly used in evaluating traditional information retrieval systems. Others are newly developed to evaluate search engine stability, an issue unique to Web information retrieval systems. An experiment was conducted to test these new measurements by applying them to a performance comparison of three commercial search engines: Google, AltaVista, and Teoma. Twenty-four subjects ranked four sets of Web pages and their rankings were used as benchmarks against which to compare search engine performance. Results show that the proposed measurements are able to distinguish search engine performance very well.

Keywords: Web search engines; Evaluation criteria; Information retrieval experiment

1. Introduction

The astonishing growth of the Web propelled the rapid development of Web search engines. However, the evaluation of these search engines has not been keeping up with the pace of their development. The significance of the evaluation is twofold: to help Web users in their choice of search engines and to inform the development of search algorithms and search engines.

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. Among the most commonly used criteria are recall and precision. Precision is the proportion of retrieved documents that are relevant, while recall is the proportion of relevant documents that

* Tel.: +1-519-661-2111x88499; fax: +1-519-661-3506.
E-mail address: lvaughan@uwo.ca (L. Vaughan).

0306-4573/$ - see front matter
doi:10.1016/S0306-4573(03)00043-8

are retrieved (Voorhees & Harman, 2001, p. 5). However, it is very difficult, if not impossible, to directly apply these measurements to the evaluation of Web information retrieval systems due to the unique nature of the Web. Variations of these measurements have been proposed and used in earlier studies as discussed in the "related studies" below. Most of these modified measures still rely on binary relevance decisions (relevant vs. non-relevant) or multi-level discrete relevance judgements (e.g. relevant, partially relevant, non-relevant). This study proposes a set of measurements that are based on a continuous ranking (from the most relevant to the least relevant) of the set of experiment documents (or Web pages). A previous study shows that human subjects are able to make relevance judgements on a continuous scale (Greisdorf & Spink, 2001). The main justification for using a continuous ranking rather than discrete relevance judgements is that retrieval results from Web search engines are typically ranked. Measurements based on ranking will therefore provide a better "match" with the system being evaluated.

Studies on the evaluation of Web search engines have proposed various measures ranging from coverage to response time (e.g. Chu & Rosenthal, 1996; Gwizdka & Chignell, 1999). However, none has recommended that performance stability be an evaluation criterion. While traditional information retrieval systems such as DIALOG provide very stable search results for a given query executed multiple times, Web search results can be very unstable due to the unique environment in which the Web search engines operate. For example, Web search engines can truncate results to improve response times during peak periods of activity. Multiple databases or multiple indexes, which are not always identical, may be used by the same search engine to respond to user queries (Mettrop & Nieuwenhuysen, 2001, pp. 641–642). It is therefore very important to include performance stability in any Web search engine evaluation. If a search engine is not stable, then the results obtained from the search engine for evaluation purposes may be just a fluke and may not represent the general performance of the search engine. Another reason for testing search engine stability is to provide information to researchers who use Web search engines to collect data, e.g. Webometrics research. These researchers need to know the stability of a search engine to gauge the reliability of the data collected. This study thus proposed a set of measurements to evaluate performance stability.

An experiment was conducted to test the measurements proposed by applying them to the comparison of three Web search engines. Four sets of Web pages corresponding to four queries were retrieved from the search engines and the ranking of these pages by each engine recorded. Twenty-four human subjects ranked these four sets of Web pages and their rankings were used as the benchmark against which the ranking results of different search engines were compared. A search engine that generated a ranking closer to the human ranking is considered better. To assess the stability of the search results, queries in the study were performed on each search engine once a week over a 10-week period.

2. Related studies

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). Only studies that used formal evaluation procedures are reviewed below.

2.1. Recall in search engine 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 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 judgement 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.

2.2. Precision in search engine evaluation

Precision is always reported in formal information retrieval experiments. However, there are variations in the way it is calculated depending on how relevance judgements were made. "TREC almost always uses binary relevance judgement—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 judgement and calculated traditional recall and precision measures at various cut-off levels. Recognizing the fact that there could a degree of relevance, many studies used multi-level rather than binary relevance judgements. For example, Chu and Rosenthal (1996) used a three-level relevance score (relevant, somewhat relevant, irrelevant) while Gordon and Pathak (1999) used a four-level relevance judgement (highly relevant, somewhat relevant, somewhat irrelevant, 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 judgement. 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 judgements. The correlation between human ranking and search engine ranking can be calculated instead of the traditional measure of precision.

2.3. Human relevance judgements

Not all search engine studies used human relevance judgement 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 judgement, which would be extremely difficult to do given the total number of items for which relevance judgements 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 judgement was used, there was a variation in who makes the judgement. TREC leaves relevance judgements 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 judgements can only be made by the individual with the original information need. The study reported in this paper used a group of 24 subjects for relevance judgement. However, the search queries did not come directly from the subjects due to the lack of such a source and the plan to randomly assign subjects to test groups. Search topics (all about Canadian university life) were carefully selected to be of interest and relevant to the subjects (Canadian graduate students) so that the subjects are knowledgeable about the topics and competent to rank the Web pages in the experiment.

2.4. Search engine stability measures

Studies that address the issue of search engine stability are largely concerned with Webometrics data collection. Search engine performance was found to be very volatile in earlier years (e.g. Rousseau, 1998/99; Snyder & Rosenbaum, 1999), although recent testing has found improvement (Thelwall, 2001; Vaughan & Hysen, 2002; Vaughan & Thelwall, 2003). Bar-Ilan conducted several studies to investigate the search engine stability problem (Bar-Ilan, 1998/99, 2000, 2002) and defined several measures to evaluate search engine functionality over time (Bar-Ilan, 2002). Bar-Ilan's measures are based on the "technical relevance" concept: a document is defined to be technically relevant if it fulfills all the conditions posed by the query. The advantage of the

"technical relevance" approach is that it can be calculated quickly and easily by a simple computer program instead of requiring human relevance judgement. Using this method, Bar-Ilan was able to test the performance of six largest search engines over a one-year period. The study reported here does not use the measure recommended by Bar-Ilan because the "technical relevance" concept is not appropriate in this study.

After testing the consistency of 13 search engines, Mettrop and Nieuwenhuysen (2001) concluded that search engine fluctuations in the result sets can no long be neglected and that search result stability should be considered as a performance measure of Internet search engines. Bar-Ilan (2002) also recommend setting up a standard suite of measures for evaluating search engine performance. However, to date, there is no standard for measuring search engine stability. In light of the absence of a standard, this study proposes a set of three stability measures that can be calculated easily.

3. Proposed measurements

Two measurements are proposed as counterparts of traditional recall and precision. In contrast to the calculations of recall and precision, which are based on binary relevance judgements, the proposed measurements are calculated based on a continuous relevance ranking (from most relevant to least relevant) by human subjects. In addition, a set of three measurements is proposed to evaluate the stability of search engine performance.

3.1. Quality of result ranking (counterpart measure of precision)

The quality of result ranking by a search engine can be measured by the correlation between engine ranking and human ranking. Specifically, the Spearman rank-order correlation coefficient can be calculated and viewed as a counterpart measure of precision. The higher the correlation coefficient, the closer the engine ranking is to the human ranking and thus the better the engine performance.

3.2. Ability to retrieve top ranked pages (counterpart measure of recall)

A modified recall is proposed which is calculated as the percent of top ranked Web pages that are retrieved, i.e. the engine's ability to retrieve top ranked pages. The following steps are taken to do this calculation. First, a search query is executed on each search engine being examined. Then, the top pages (e.g. top 10 pages) retrieved by each engine are taken and are merged into a single set. Human subjects will rank the merged set of pages according to their relevance to the query. It is likely that non-relevant pages are retrieved and ranked among the top 10 by one or more of the search engines. However, it is safe to assume that these non-relevant pages do not form the majority of the pages in the set and that they will be ranked lower by human subjects. Assuming that these non-relevant pages constitute 25% of the pages in the set and that they are ranked by humans in the bottom 25%, then the top 75% pages can be considered as relevant pages that should be retrieved. The percent of these relevant pages that were retrieved by each

search engine can be calculated (a kind of modified recall). We can change the cut-off point from top 75% to top 50% (i.e. consider the top 50% pages as relevant pages) or to any other numbers.

3.3. Stability measurements

A set of three measurements is proposed 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 20 pages retrieved need to be 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 20 pages is suggested because it will be extremely time consuming to compare each page in the entire retrieved set, which is typically very large. Another justification for restricting to the top 20 pages is that few, if any, users go through all of the pages retrieved (Silverstein, Henzinger, Marais, & Moricz, 1999; Spink, Jansen, Wolfram, & Saracevic, 2002). The recommendation of "top 20 pages" can be changed to "top 15 pages" or "top 10 pages" to save labour in the study if needed.

4. Experiment to test the proposed measurements

An experiment was carried out to test the proposed measurements by applying them in the comparison of three search engines. The design of the experiment is detailed below.

4.1. Selection of search engines

The three search engines selected were: Google (www.google.com), AltaVista (www.altavista.com), and Teoma (www.teoma.com). Google was selected because it was the largest publicly available search engine at the time of the study (Notess, 2002). Its PageRank algorithm pioneered the new generation of link-based ranking algorithm, which is very different from the term frequency-based vector space model (Salton & McGill, 1983). AltaVista is one of the oldest Web search engines and was among the largest search engines by database size at the time of the study (Notess, 2002). Another reason that AltaVista was included in the study is that it is the most commonly used search engine for data collection for Webometrics studies (e.g. Vaughan & Thelwall, 2003). Teoma, founded in 2000 (Teoma, 2002a), is the youngest of the three. It uses a unique ranking algorithm, Subject-Specific Popularity, to determine the authority or quality of a site's content. Subject-Specific Popularity ranks a site based on the number of same-subject pages that link it, not the general popularity that Google uses, to determine a site's level of authority (Teoma, 2002b).

4.2. Search queries and Web pages retrieved

Queries in the study were designed to test various search features including single word search, phrase search, and a combination of the two using a Boolean operator. The four search topics and their corresponding search queries were:

1. Topic: Society of Graduate Studies at the University of Western Ontario (acronym SOGS).
   Search query: SOGS.
   Type of query tested: single word search.
   Search restricted to domain: www.uwo.ca.
   Query referred to later in this paper as **SOGS.**
2. Topic: Ontario Graduate Scholarship in Science and Technology.
   Search query "Ontario Graduate Scholarship in Science and Technology".
   Type of query tested: phrase search.
   Search restricted to domain: .ca (note: .ca is the top level domain name for Canada).
   Query referred to later in this paper as **OGS**.
3. Topic: ombudsperson office at the University of Western Ontario.
   Search query: ombudsperson AND office.
   Type of query tested: two word searches connected by a Boolean "AND".
   Search restricted to domain: uwo.ca.
   Query referred to later in this paper as **ombudsperson.**
4. Topic: admission requirements for the MBA program at the University of Toronto.
   Search query: "admission requirements" AND MBA.
   Type of query tested: a phrase search and a word search connected by a Boolean "AND".
   Search restricted to domain: utoronto.ca.
   Query referred to later in this paper as **MBA.**

Different search engines have different search functions and syntaxes. Previous studies (Gordon & Pathak, 1999; Hawking et al., 2001) recommended that the most effective combination of specific features of each search engine be exploited. In other words, the queries submitted to the engines need not be the same and should be designed to take advantage of the search engine. Based on this principle, the following adjustments were made to the search queries. First of all, different syntaxes of Boolean operators and domain restrictions were considered and the appropriate one used for each engine. For example, the queries for the first topic were: "SOGS AND host: www.uwo.ca" for AltaVista, "SOGS site:www.uwo.ca" for Google, and "SOGS www.uwo.ca" for Teoma. Second, the query on "ombudsperson" was chosen to be a word search rather than a phrase search because the possible wordings of "office of ombudsperson" and "ombudsperson office".

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). As a result of the merge, the number of pages in the sets **MBA**, **OGS**, **ombudsperson**, and **SOGS** were 21, 26, 16, and 18 respectively. 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). The justification for restricting attention to the top 10 pages is that users typically do not go through the entire search results but rather visit only the top pages retrieved. A study by Silverstein et al. (1999) revealed that 85.2% AltaVista users viewed only top 10 pages. Spink et al. (2002) reported that the trend of viewing fewer pages of search results is going up.

Each search engine's ranking for each page in the final sets was recorded. These rankings were later compared with the human rankings to determine which engine's ranking was better. It should be noted that not every page in the final merged set was ranked among the top 10 by each search engine. For example, a page may be ranked third by search engine A but ranked 100th by search B. Each search engine's ranking for each page was determined by going through the entire search results, not just the top 10 pages. If a particular search engine did not retrieve a page, then that page was marked as "not retrieved" by that engine.

4.3. Human ranking of the Web pages

Subjects in the study were graduate students enrolled in the Information Retrieval course in the 2002 summer term at the Faculty of Information and Media Studies, University of Western Ontario, Canada. An assignment on search algorithms and search engines required students to rank a set of Web pages and then compare their own ranking with those generated by various search algorithms. The Web pages used in this exercise were those used in the current study.

The 24 students in the course were divided randomly into four groups of six people each. Each group was given a set of Web pages (i.e. the four groups were given the four sets of experiment pages). Each student was asked to independently rank the pages in the way that he/she thinks they should be ranked in a search output and to write down the criteria used in the ranking. The group then met and discussed their ranking as well as ranking criteria. This was done out of concern for individual variations in relevance assessment found in earlier studies (e.g. Harter, 1996; Maglaughlin & Sonnenwald, 2002). The purpose of this group meeting was to improve the ranking quality through the group consensus method that was shown to be beneficial in an earlier study (Zhang, 2002).

Students could change their ranking based on the group discussion. Individual ranking results were aggregated by taking a group average, a process that was expected to reduce the effects of the possible unusual rankings by individual students. The group average was considered to represent a human ranking against which ranking results from search engines could be compared. The rankings of these pages by search engines were not revealed to subjects before their ranking exercise to avoid possible bias. The reliability of the student ranking data is shown by the fact that individual student's rankings, although they do not match perfectly with each other, are significantly correlated with each other in most cases, even before the group meeting. Due to the ethical requirement of voluntarily participation, students were given the option of not allowing their ranking data to be used for the study. All students in the course gave the permission to use their data.

4.4. Search engine stability test

Each query was searched on each search engine once a week through a 10-week period. The time and the day to conduct the search were fixed: every Wednesday from 10 pm on, from May 1, 2002 to July 3, 2002 inclusive. In each week's data collection, the number of pages retrieved by each search engine for each query was recorded. The top 20 pages retrieved were compared with those from the previous week to determine (1) if the pages retrieved were the same; (2) if the ranking of the pages was the same.

5. Experiment results

5.1. Quality of result ranking

The quality of result ranking was measured by the correlation between search engine ranking and human ranking. The Spearman correlation coefficient was calculated for each query and for each search engine. The results are summarized in Table 1. A higher correlation coefficient indicates better quality of ranking by the search engine. Coefficients that are statistically significant at 0.05 level are indicated by a "*" sign beside them. Each search engine's performance over the four queries was summarized by the arithmetic average of the four correlation coefficients and presented in the last column of Table 1.

It is clear from Table 1 that Google is the best in result ranking. Its correlations for all four queries were statistically significant and the average of 0.72 is remarkably high, much higher than that of the other two engines. AltaVista achieved significant correlations for three out of the four queries, a respectable performance. Teoma performed poorly with no significant correlation with human ranking in any of the four queries. In fact, its negative correlation for query **OGS** shows that this ranking bears no resemblance with that of the human ranking. An examination of Teoma's search results for the query **OGS** revealed the underlying cause of this poor performance: Teoma's phrase search did not work properly. The query **OGS** was searched as a phrase "Ontario Graduate Scholarship in Science and Technology". However, Teoma seemed to interpret the phrase search as a word search connected by Boolean "AND". For example, one of the top10 hits retrieved reads: "General Information regarding **graduate** fellowships **and** scholarships is posted on the **Graduate** Studies bullet**in** board **in** the Concourse of ..." Teoma boldfaced words used in the search. Apparently, not only were the words "and", and "in" used in the search but the word "bulletin" was also used because of the "in" at the end of the word! Oddly, the word "scholarships" was not boldfaced (i.e. not used in the search) although the word "scholarship" was in the query.

Table 1
Correlation between human ranking and search engine ranking

	MBA	OGS	Ombudsperson	SOGS	Average
AltaVista	0.68*	0.07	0.73*	0.47*	0.49
Google	0.66*	0.66*	0.84*	0.8*	0.72
Teoma	0.54	−0.42	0.5	0.16	0.19

Table 2
Percent of top ranked pages retrieved (top 75% pages as relevant pages)

	MBA (%)	OGS (%)	Ombudsperson (%)	SOGS (%)	Average (%)
AltaVista	50	80	100	100	82.5
Google	87.5	80	100	100	91.9
Teoma	68.8	70	58.3	100	74.3

Table 3
Percent of top ranked pages retrieved (top 50% pages as relevant pages)

	MBA (%)	OGS (%)	Ombudsperson (%)	SOGS (%)	Average (%)
AltaVista	63.6	69.2	100	100	83.2
Google	90.9	76.9	100	100	92
Teoma	63.6	69.2	50	100	70.7

5.2. Ability to retrieve top ranked pages

Assuming that the top 75% pages ranked by human are relevant pages that should be retrieved, we can then calculate the percent of these relevant pages that were retrieved by each search engine as a counterpart measure of the traditional recall. Table 2 shows the results of this calculation. The last column of Table 2 is the average over the four queries. Alternatively, we can change the cut-off level from top 75% to top 50% (i.e. consider the top 50% pages as relevant pages) and calculate this "modified recall" again as shown in Table 3.

Results from the two tables are consistent: Google has the highest success rate in retrieving relevant pages, followed by AltaVista. Teoma again performed worst among the three. It should be noted that the figures in Tables 2 and 3 cannot be compared directly with the recall figures obtained in traditional information retrieval experiments because the two are calculated in very different ways. However, it is valid to compare the three search engines here using these numbers. It is interesting to note that the figures in the "average" column of Tables 2 and 3 are similar, which suggests that the "ability to retrieve top ranked pages" are about the same regardless what cut-off level (top 50% or top 75%) is used. In other words, the relative comparisons of the three engines remain the same regardless of the cut-off level used.

5.3. Stability comparison

The stability of search engine performance was first examined by looking at the stability of the number of pages retrieved over the 10-week period. The comparisons of the three engines in this regard are shown from Figs. 1–4 for the four queries respectively. The four Figures present the same pattern and thus same conclusions: (1) Google consistently retrieved more pages than AltaVista and Teoma; and (2) Google remained very stable over the 10-week period without any sudden increase or decrease in the number of pages retrieved. In contrast, AltaVista had a sudden surge, almost double the number of pages retrieved for all four queries in week 7, perhaps the result of a major increase in database size. Teoma's stability is between that of Google and AltaVista.

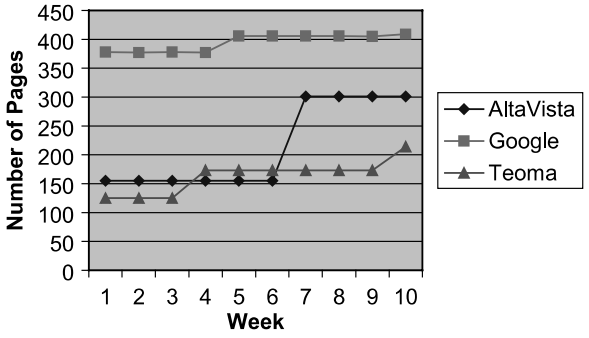

Fig. 1. Search results for query "MBA".

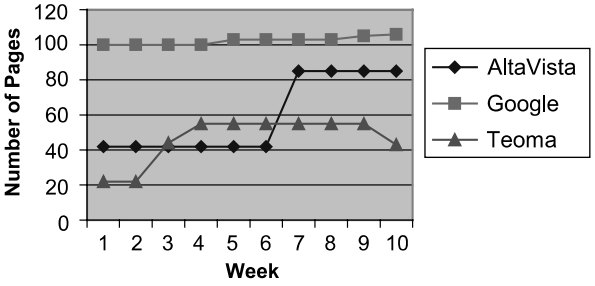

Fig. 2. Search results for query "ombudsperson".

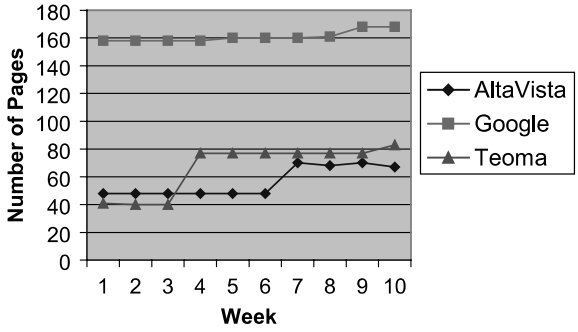

Fig. 3. Search results for query "OGS".

The second measure of search engine stability is to compare the top 20 pages retrieved by a particular search engine for a particular search query in two consecutive weeks and record the number of pages that are the same. Table 4 shows the average over the 10-week period for each query by each search engine. The last column of Table 4 is the average over the four queries. Google again stands out as the best, having an average of 98.5% (19.7/20) pages that are the same in two consecutive weeks. Teoma is the second, followed by AltaVista.

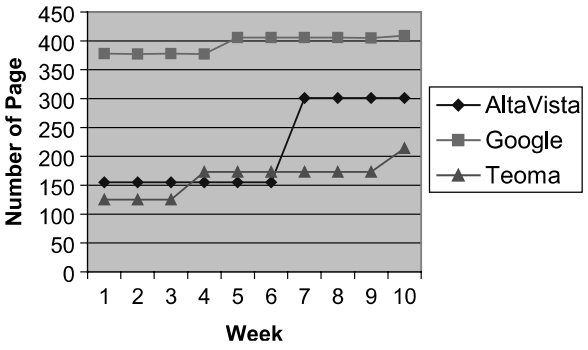

Fig. 4. Search results for query "SOGS".

Table 4
Same pages retrieved in two consecutive weeks—average over 10 weeks for the top 20 pages

	MBA	OGS	Ombudsperson	SOGS	Average
AltaVista	18.2	16.9	17.4	19.2	17.9
Google	19.7	19.3	20	19.9	19.7
Teoma	18	18.4	18.6	18.7	18.4

Table 5
Number of pages that remain the same ranking order—average over 10 weeks for the top 20 pages

	MBA	OGS	Ombudsperson	SOGS	Average
AltaVista	16.8	15.1	17.1	18.3	16.8
Google	18.9	18.6	19.8	19.7	19.3
Teoma	17.9	17.9	18.2	18.2	18.1

The third measure of search engine stability is to compare the top 20 pages retrieved by a particular search engine for a particular search query in two consecutive weeks and record the number of pages that remained in the same relative ranking order. Table 5 presents the average over the 10-week period for each query by each search engine. The last column of Table 5 is the average over the four queries. Google is again the most stable followed by Teoma and AltaVista.

It should be noted that AltaVista, the search engine that has been used most frequently in Webometrics studies due to its advanced Boolean search capabilities, measured worst in stability among the three search engines. Although no definitive conclusions can be reached about AltaVista's stability from this study due to the limitations of data collected in a 10-week period, Webometrics researchers should be aware of this study result and pay close attention to the issue.

6. Discussion and conclusions

Two measurements are proposed as counterparts of traditional recall and precision: the quality of result ranking and the ability to retrieve top ranked pages. The main difference between these

measurements and those used in earlier studies is that these new measures are based on a continuous ranking of test documents (ranked from the most relevant to the least relevant) rather than the discrete relevance judgements (e.g. relevant, partial relevant, irrelevant) used in previous studies. It is argued that these new measures based on ranking are more effective in evaluating Web search engines, most of which provide ranked results. It should be acknowledged that these new measures have disadvantages. The ranking poses more cognitive burdens on human subjects than the discrete relevance judgement does. Human may not be able to reliably rank a large set of documents, which restricts the size of experiment set. Nevertheless, it is contended that the better "match" between these ranking-based evaluation criteria and the "ranked" nature of Web search results shows the merits of these new measures and justifies their use.

An experiment was carried out to test these new measurements by applying them to the comparison of three search engines. The experiment results show that these measurements are able to distinguish search engine performance and Google performed best. The better performance of Google under these measurements echoes the conclusion reached in other studies (Hawking et al., 2001; Thelwall, 2002) where different measurements were used. Google's relative superiority is also reflected in the fact it is the most popular search engine on the Web (AltaVista ranked 9th among the 12 compared) (Sullivan, 2002). This evidence provides some assurance of the validity of the measurements proposed in this study.

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 four 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. This is, however, beyond the resources of an individual researcher and requires the effort of groups of researchers such as those seen in TREC. A full-scale testing of the measurements proposed here is possible in a TREC type of environment where human relevance judgements have always been used.

The study also proposed to include in the evaluation criteria measurements of performance stability and proposed a set of three measurements to examine the stability of search engine performance. The stability criteria were not needed in evaluating traditional information retrieval systems such as DIALOG because the same search results for a given query can be expected unless additional items are added to the collection and the indexes are updated accordingly. This kind of performance consistency is not always achieved in Web search engines. The same query can retrieve drastically different number of pages over a period of a few days or even within the same day (Snyder & Rosenbaum, 1999). This is not likely to be the result of updates in the index. The fact that the same page can appear, disappear, and then reappear in a sequence of searches (Bar-Ilan, 2002) confirms this assertion. This kind of performance instability is obviously an undesirable feature and the proposed measurements are meant to assess the extent of this problem. However, the proposed measurements should not be interpreted to mean that the same search results should always be expected over time for a given query. Obviously, when a search engine finds new pages and updates its index, the search results should reflect the change. Regular updating of the index is a positive feature and stagnation is a negative one. Therefore, the underlying assumption of the proposed measurements is that there is no update in the index. Bearing this

assumption in mind, we need to select search queries carefully so that frequent updating of pages on the query topics is not likely to occur.

Acknowledgements

I am very grateful to all the students who participated in the study by giving permission for me to use their ranking data. The study would have been impossible without their support. I also thank the two anonymous referees for their very helpful comments and suggestions.

References

Bar-Ilan, J. (1998/99). Search engine results over time—a case study on search engine stability. *Cybermetrics, 2/3*(1), Retrieved 15 December 2002 from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html.

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

Bar-Ilan, J. (2002). Methods for assessing search engine performance over time. *Journal of the American 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.

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, Retrieved 18 December 2002 from http://citeseer.nj.nec.com/brin98anatomy.html.

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.

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

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.

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

Greisdorf, H., & Spink, A. (2001). Median measure: an approach to IR Systems evaluation. *Information Processing & Management, 37*(6), 843–857.

Gwizdka, J., & Chignell, M. (1999). Towards information retrieval measures for evaluation of Web search engines. Retrieved 17 December 2002 from http://www.imedia.mie.utoronto.ca/~jacekg/pubs/webIR_eval1_99.pdf.

Harter, S. (1996). Variation in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science, 47*(1), 37–49.

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

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

Maglaughlin, K. L., & Sonnenwald, D. H. (2002). User perspectives on relevance criteria: a comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the American Society for Information Science and Technology, 53*(5), 327–342.

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

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

Notess, G. R. (2002). Search engine statistics: relative size showdown. Retrieved 2 September 2002 from http://www.searchengineshowdown.com/stats/size.shtml.

Rousseau, R. (1998/99). Daily time series of common single word searches in Alta Vista and Northern Light. *Cybermetrics, 2/3*(1), Retrieved 15 December 2002 from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

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

Schlichting, C., & Nilsen, E. (1996). Signal detection analysis of WWW search engines. Presented at *Microsoft's Designing for the Web: Empirical Studies Conference, October 1996*. Retrieved 17 December 2002 from http://www.microsoft.com/usability/webconf/schlichting/schlichting.htm.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. *SIGIR Forum, 33*(1), 6–12.

Snyder, H., & Rosenbaum, H. (1999). Can search engines be used as tools for Web-link analysis? A critical view. *Journal of Documentation, 55*(4), 375–384.

Spink, A., Jansen, B., Wolfram, D., & Saracevic, T. (2002). From E-Sex to E-Commerce: Web search changes. *IEEE Computer, 35*(2), 133–135.

Su, L. T., Chen, H., & Dong, X. (1998). Evaluation of Web-based search engines from the end-user's perspective: a pilot study. In *Proceedings of the 61st annual meeting of the American Society for Information Science* (pp. 348–361).

Sullivan, D. (2002). Nielsen//NetRatings: search engine ratings. Retrieved 23 December 2002 from http://www.searchenginewatch.com/reports/netratings.html.

Teoma (2002a). Teoma's history. Retrieved 23 December 2002 from http://sp.teoma.com/docs/teoma/about/developmentteamhistory.html.

Teoma (2002b). Search with authority: The Teoma difference. Retrieved 23 December 2002 from http://sp.teoma.com/docs/teoma/about/searchwithauthority.html.

Thelwall, M. (2001). The responsiveness of search engine indexes. *Cybermetrics, 5*(1), Retrieved 17 December 2002 from http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html.

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

Vaughan, L., & Hysen, K. (2002). Relationship between links to journal Web sites and impact factors. *Aslib Proceedings: New Information Perspectives, 54*(6), 356–361.

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.

Voorhees, E. M., & Harman, D. (2001). Overview of TREC 2001. *NIST Special Publication 500-250: The 10th text retrieval conference (TREC 2001)* (pp. 1–15). Retrieved 17 December 2002 from http://trec.nist.gov/pubs/trec10/papers/overview_10.pdf.

Zhang, X. (2002). Collaborative relevance judgment: A group consensus method for evaluating user performance. *Journal of the American Society for Information Science and Technology, 53*(3), 220–231.