Dynamics of Search Engine Rankings – A Case Study

Judit Bar-Ilan
The Hebrew University of Jerusalem and Bar-Ilan University Israel
judit@cc.huji.ac.il

Mark Levene and Mazlita Mat-Hassan
School of Computer Science and Information Systems
Birkbeck , University of London
{mark, azy}@dcs.bbk.ac.uk

Abstract

The objective of this study was to characterize the changes in the rankings of the top-n results of major search engines over time and to compare the rankings between these engines. We considered only the top-ten results, since users usually inspect only the first page returned by the search engine, which normally contains ten results. In particular, we compare rankings of the top ten results of the search engines Google and AlltheWeb on identical queries over a period of three weeks. The experiment was repeated twice, in October 2003 and in January 2004 in order to assess changes to the top ten results of some of the queries during a three months period. Results show that the rankings of AlltheWeb were highly stable over each period, while the rankings of Google underwent constant yet minor changes, with occasional major ones. Changes over time can be explained by the dynamic nature of the Web or by fluctuations in the search engines' indexes (especially when frequent switches in the rankings are observed). The top ten results of the two search engines have surprisingly low overlap. With such small overlap (occasionally only a single URL) the task of comparing the rankings of the two engines becomes extremely challenging, and additional measures are needed to assess rankings in such situations.

Introduction

The Web is growing continuously; new pages are published on the Web every day. However it is not enough to publish a Web page – this page must also be locatable. Currently the primary tools for locating information on the Web are the search engines, and by far the most popular search engine is Google (Nielsen/NetRatings, 2003; Sullivan & Sherman, 2004).

Google reportedly covers over 4.2 billion pages as of mid-February 2004 (Google, 2004; Price, 2004), a considerable jump from over 3.3 billion as reported from August 2003 and until mid-February 2004. Some of the pages indexed by Google are not from the traditional "publicly indexable Web" (Lawrence & Giles, 1999), for example records from OCLC's WorldCat (Quint, 2003). Currently the second largest search engine in terms of the reported number of indexed pages is AlltheWeb with over 3.1 billion pages (AlltheWeb, 2004). At the time of our data collection, the two search engines were of similar size. There are no recent studies on the coverage of Web search engines, but the 1999 study of Lawrence and Giles found that the, then largest search engine (NorthernLight), covered only about 16% of the Web. Today, authors of Web pages can influence the inclusion of their pages through the paid-inclusion services. AlltheWeb has a paid-inclusion service, and even though Google doesn't, one's chances of being crawled are increased if the pages appear in major directories (which do have paid-inclusion services) (Sullivan, 2003a).

However, it is not enough to be included in the index of a search engine, placement is also crucial, since most Web users do not browse beyond the first ten or twenty results (Silverstein et al., 1999; Spink et al., 2002). Paid inclusion is not supposed to influence the placement of the page. The SEOs (Search Engine Optimizers) offer their services to increase the ranking of

your pages on certain queries (see for example Search Engine Optimization, Inc, http://www.seoinc.com/) – Google (Google, 2003a) warns against careless use of such services. Thus it is clear to all that the top ten results retrieved on a given query have the best chance of being visited by Web users. This was the main motivation for the research we present herein, in addition to examining the changes over time in the top ten results for a set of queries of the currently two largest search engines, Google and AlltheWeb. In parallel to this line of enquiry, we also studied the similarity (or rather non-similarity) between the top ten results of these two tools.

For this study, we could not analyze the ranking algorithms of the search engines, since these are kept secret, both because of the competition between the different tools and in order to avoid misuse of the knowledge of these algorithms by users who want to be placed high on specific queries. For example, Google is willing to disclose only that its ranking algorithm involves more than 100 factors, but "due to the nature of our business and our interest in protecting the integrity of our search results, this is the only information we make available to the public about our ranking system" (Google, 2003b). Thus we had to use empirical methods to study the differences in the ranking algorithms and the influence of time on the rankings of search engines.

The usual method of evaluating rankings is through human judgment. In an early study by Su et al. (1998), users were asked to choose and rank the five most relevant items from the first twenty results retrieved for their queries. In their study, Lycos performed better on this criteria than the other three examined search engines. Hawking et al. (1999) compared precision at 20 of five commercial search engines with precision at 20 of six TREC systems. The results for the commercial engines were retrieved from their own databases, while the TREC engines' results came from an 18.5 million pages test collection of Web pages. Findings showed that the TREC systems outperformed the Web search engines, and the authors concluded that "the standard of document rankings produced by public Web search engines is by no means state-of-the-art." On the other hand, Singhal and Kaszkiel (2001) compared a well-performing TREC system with four Web search engines and found that "for finding the web page/site of an entity, commercial web search engines are notably better than a state-of-the-art TREC algorithm." They were looking for home pages of the entity and evaluated the search tool by the rank of the URL in the search results that pointed to the desired site. In Fall 1999, Hawking et al. (2001) evaluated the effectiveness of twenty public Web search engines on 54 queries. One of the measures used was the reciprocal rank of the first relevant document – a measure closely related to ranking. The results showed significant differences between the search engines and high intercorrelation between the measures. Chowdhury and Soboroff (2002) also evaluated search effectiveness based on the reciprocal rank – this time of the URL of a known item.

Evaluations based on human judgments are unavoidably subjective. Voorhees (2000) examined this issue, and found very high correlations among the rankings of the systems produced by different relevance judgment sets. The paper considers rankings of the different systems and not rankings within the search results, and despite the fact that the agreement on the ranking performance of the search tools was high, the mean overlap between the relevance judgments on individual documents of two judges was below 50% (binary relevance judgments were made). Soboroff et al. (2001) based on the finding that differences in human judgments of relevance do not affect the relative evaluated performance of the different systems, proposed a ranking system based on randomly selecting "pseudo-relevant" documents. In a recent study, Vaughan (to appear) compared human rankings of 24 participants with those of three large commercial search engines, Google, AltaVista and Teoma on four search topics. The highest average correlation between the human-based rankings and the rankings of the search engines was for Google, where the average correlation was 0.72. The average correlation for AltaVista was 0.49.

Fagin et al. (2003) proposed a method for comparing the top-*k* results retrieved by different search engines. One of the applications of the metrics proposed by them was comparing the rankings of the top 50 results of seven public search tools (AltaVista, Lycos, AlltheWeb, HotBot, NorthernLight, AOLSearch and MSNSearch - some of them received their results from the same source, e.g., Lycos and AlltheWeb) on 750 queries. The basic idea of their method was to assign some reasonable, virtual placement to documents that appear in one of the lists but not in the other. The resulting measures were proven to be metrics, which is a major point they stress in their paper.

The studies we have mentioned concentrate on comparing the search results of several engines at one point in time. In contrast, this study examines the temporal changes in search results over a period of time within a single engine and between different engines. In particular, we concentrate on the results of two of the largest search engines, Google and AlltheWeb using three different measures described below.

Methodology

Data Collection

The data for this study was collected during two, approximately three weeks long time periods, the first during October 2003 and the second during January 2004. The data collection for the first period was a course assignment at Birbeck, University of London. Each student was required to choose a query from a list of ten queries and also to choose an additional query of his/her own liking. These two queries were to be submitted to Google (google.com) and AlltheWeb (alltheweb.com) twice a day (morning and evening) during a period of three weeks. The students were to record the ranked list of the top ten retrieved URLs for each search point. Overall, 34 different queries were tracked by twenty-seven students (some of the queries were tracked by more than one student). The set of all queries that were processed with the numbering assigned to them appear in Table 1. For the first period queries q01-q05 were analyzed.

The process was repeated at the beginning January 2004. We picked 10 queries from the list of 34 queries. This time we queried Google.com, Google.co.uk, Google.co.il and Alltheweb in order to assess the differences between the different Google sites as well. In this experiment, at each data collection point all the searches were carried out within a 20-minute timeframe. The reason for rerunning the searches was to study the effect of time on the top ten results. Between the two parts of the experiment, Google most likely introduced a major change into its ranking algorithm (called the "Florida Google Dance" - (Sullivan, 2003b)), and we were interested to study the effects of this change. For the second period queries q01-q10 were analyzed. The search terms were not submitted as phrases at either stage.

Query ID	Query
q01	Modern architecture
q02	Web data mining
q03	world rugby
q04	Web personalization
q05	Human Cloning
q06	Internet security
q07	Organic food
q08	Snowboarding
q09	dna evidence
q10	internet advertising techniques

Table 1: The queries

The Measures

We used three measures in order to assess the changes over time in the rankings of the search engines and to compare the results of Google and AlltheWeb. The first and simplest measure is simply the size of the overlap between two top ten lists.

The second measure was Spearman's rho. Spearman's rho is applied to two rankings of the same set, thus if the size of the set is *N*, all the rankings must be between 1 and *N* (ties are allowed). Since the top ten results retrieved by two search engines on a given query, or retrieved by the same engine on two consecutive days are not necessarily identical, the two lists must be transformed before Spearman's rho can be computed. First the non-overlapping URLs were eliminated from both lists, and then the remaining lists were reranked, each URL was given its relative rank in the set of remaining URLs in each list. After these transformations Spearman's rho could be computed:

$$r = 1 - \frac{6\sum d_i^2}{(n^2 - 1)n}$$

where d_i is the difference between the ranking of URL_i in the two lists. The value of r is between -1 and 1, where -1 indicates that the two lists have opposite rankings, and 1 indicates perfect correlation. Note that Sperman's rho is based on the reranked lists, and thus for example if the original ranks of the URLs that appear in both lists (the overlapping pairs) are (1,8), (2,9) and (3,10), the reranked pairs will be (1,1), (2,2) and (3,3) and the value of Spearman's rho will be 1 (perfect correlation).

The third measure utilized by us was one of the metrics introduced by Fagin et al. (2003). It is relatively easy to compare two rankings of the same list of items – for this well-known statistical measures such as Kendall's tau or Spearman's rho can be easily utilized. The problem arises when the two search engines that are being compared rank non-identical sets of documents. To cover this case (which is the usual case when comparing top-*k* lists created by different search engines), Fagin et al. (2003) extended the previously mentioned metrics. Here we discuss only the extension of Spearman's footrule (a variant of Spearman's rho, which is unlike Spearman's rho is a metric), but the extensions of Kendall's tau are shown in the paper to be equivalent to the extension of Spearman's footrule. A major point in their method was to develop measures that are either metrics or "near" metrics. Spearman's footrule, is the L_1 distance between two permutations (where the rankings on identical sets can be viewed as permutations): $F(\sigma_1, \sigma_2) = \sum |\sigma_1(i) - \sigma_2(i)|$. This metric is extended for the case where the two lists are not identical, to documents appearing in one of the lists but not in the other an arbitrary placement (which is larger than the length of the list) is assigned in the second list – when comparing lists of length *k* this placement can be *k+1* for all the documents not appearing in the list. The rationale for this extension is that the ranking of those documents must be *k+1* or higher – Fagin et al. do not take into account the possibility that those documents are not indexed at all by the other search engine. The extended metric becomes:

$$F^{(k+1)}(\tau_1, \tau_2) = 2(k - z)(k + 1) + \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| - \sum_{i \in S} \tau_1(i) - \sum_{i \in T} \tau_2(i)$$

where *Z* is the set of overlapping documents, and *z* is the size of *Z*, *S* is the set of documents that are only in the first list and *T* is the set of documents that appear in the second list only. A problem with the measures proposed by Fagin et al. is that when the two lists have little in common, the non-common documents have a major effect on the measure. Our experiments show that usually the overlap between the top ten results of two search engines for an identical query is very small, and the non-overlapping elements have a major effect.

$F^{(k+1)}$ was normalized by Fagin et al. so that the values lie between 0 and 1. For k=10 the normalization factor is 110. Since $F^{(k+1)}$ is a distance measure, the smaller the value the more

similar are the two lists, however for Spearman's rho the more similar the two lists are, the value of the measure is nearer to 1. In order to be able to have some comparison between the two measures, we computed

$$G^{(k+1)} = 1 - \frac{F^{(k+1)}}{\max F^{(k+1)}}$$

which we refer to as the G metric.

Data analysis

For a given search engine and a given query we computed these measures on the results for consecutive data collection points. When comparing two search engines we computed the measures on the top ten results retrieved by both engines on the given data collection point. The two periods were compared on five queries - here we calculated the overlap between the two periods and assessed the changes in the rankings of the overlapping elements based on the average rankings.

Results and Discussion

A Single Engine over Time

AlltheWeb was very stable during both phases on all queries, as can be seen in Table 2. There were almost no changes either in the set of URLs retrieved or in the relative placement of these URLs in the top ten results. Some of the queries were monitored by several students, thus the number of data comparisons (comparing the results of consecutive data collection points) was high, For each query we present the total number of URLs identified during the period, the average and minimum number of URLs that were retrieved at both of the two consecutive data collection points (overlap). The maximum overlap was 10 for each of the queries, an overlap of 10 was rather frequent, thus we computed the percentage of the comparisons where the set of URLs was not identical in both of the points that were compared (% of points with overlap less than 10). In addition, Table 1 displays the percentage of comparisons where the relative ranking of the overlapping URLs changed and the minimal values of Spearman's rho and of G (the maximal values where 1 in all cases). Finally, in order to assess the changes in the top-ten URLs over a longer period of time, we also present the number of URLs that were retrieved in both the first and the last data collection points.

query	# days monitored	# data comparisons	# URLs identified during period	average overlap	min overlap	% of points overlap less than 10	% of points where relative ranking changed	min Spearman	min G	overlap between first and last point
q01	12	20	10	10	10	0%	0%	1	1	10
q02	17	34	11	9.91	9	9%	0%	1	1	10
q03	26	109	12	9.86	9	14%	2%	0.9	0.8	10
q04	24	100	15	9.8	9	20%	0%	1	0.873	8
q05	21	41	10	10	10	0%	0%	1	1	10

Table 2. AlltheWeb – first period

When considering the data for Google we see somewhat larger variability, but still the changes between two consecutive data points are rather small. Note that for the query number 3 (world rugby), there were frequent changes in the placement of the top ten URLs.

query	# days monitored	# data comparisons	# URLs identified during period	average overlap	min overlap	% of points overlap less than 10	% of points where relative ranking changed	min Spearman	min G	overlap between first and last point
q01	12	20	11	9.95	9	5%	10%	0.95	0.891	9
q02	17	34	12	9.88	9	12%	3%	0.983	0.933	9
q03	26	109	14	9.86	8	10%	35%	0.548	0.8	8
q04	24	100	14	9.36	7	57%	0%	1	0.691	6
q05	21	41	10	10	10	0%	54%	0.891	0.927	10

Table 3. Google.com – first period

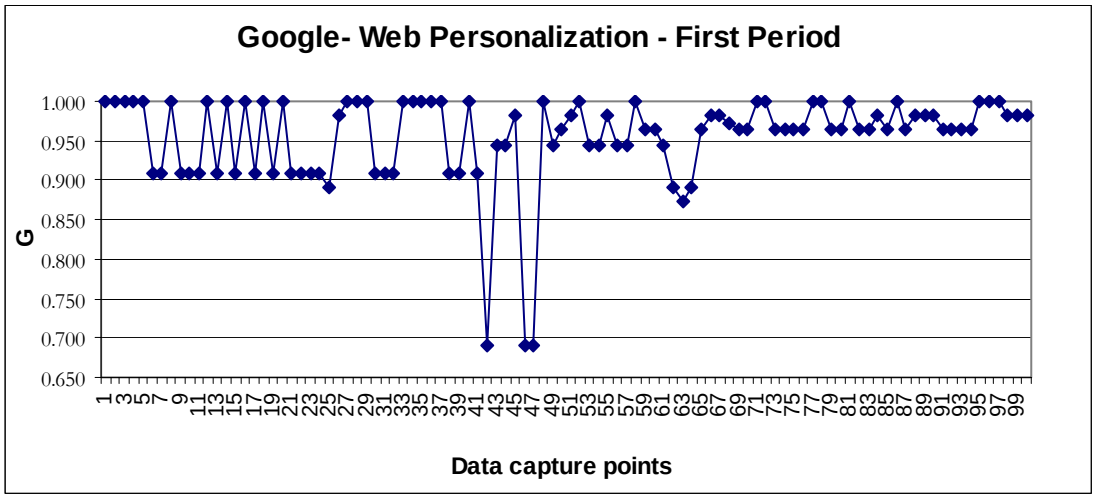

Figure 1: Time series of G metric for query, web personalization, submitted to Google.com

Figures 1 and 2 present time-series for query 4, web personalization. The x-axis for both graphs shows consecutive time-ordered data capture points. In Figure 1 we see that, during the observed period, the G metric fluctuates mainly between 0.9 and 1.0, apart from a significant drop to 0.7 during three data capture points during the middle of the period. This is due to the decrease in the size of the overlap (from 9 to 7) and changes in the ranking of the top-ten URLs observed.

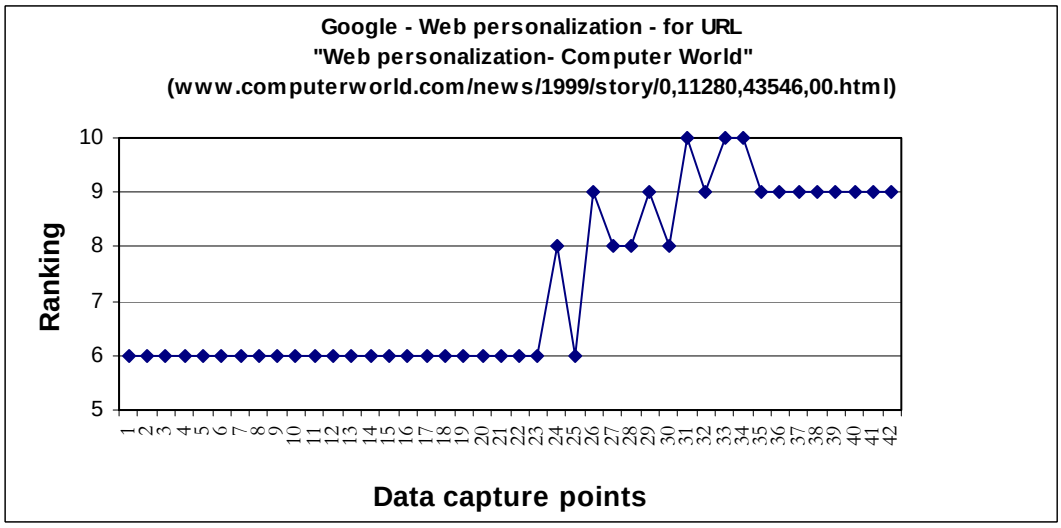

Figure 2: Time series of Google ranking for "Web personalization-Computer World" page

Figure 2 illustrates the change in Google's ranking of one web page entitled "Web personalization – Computer World", which contains an article from the Computer World website. The ranking of this page was stable at position 6, for the first twenty-three data point observed. The ranking then fluctuates between positions 8, 9 and 10 from data capture points 25 to 35. It is interesting to observe that, during this period, the rank of this Web page changed twice a day, in the morning and the evening periods. The page then settled at position 9 and then disappeared completely from the top-ten result list, three days before the end of the observed period.

Similar analysis was carried out for the queries during the second period. The results appear in Tables 4 and 5. Also during the second period the results and the rankings of AlltheWeb were highly stable. Google.com exhibited considerable variability, even though the average overlap was above 9 for all ten queries. Unlike AlltheWeb, quite often the relative placements of the URLs changed.

Perhaps the most interesting case for Google.com was query 10 (internet advertising techniques), where all except two of the previous hits were replaced by completely new ones (and the relative rankings of the two remaining URLs were swapped, and from this point on the search engine presented this new set of results. This was not accidental, the same behavior was observed on Google.co.uk and Google.co.il as well. We do not display the results for Google.co.uk and Google.co.il here, since the descriptive statistics are very similar, even though there are slight differences between the result sets. We shall discuss this point more extensively when we compare the results of the different engines.

query	# days monitored	# comparisons	# URLs identified during period	average overlap	min overlap	% of points overlap less than 10	% of points where relative ranking changed	min Spearman	min G	overlap between first and last point
q01	22	44	11	9.97	9	2%	0%	1	0.84	9
q02	22	44	11	9.97	9	2%	0%	1	0.945	9
q03	22	44	11	9.97	9	2%	0%	1	0.818	9
q04	22	44	13	9.76	8	21%	0%	1	0.89	8
q05	22	44	10	10	10	0%	0%	1	1	10
q06	22	44	10	10	10	0%	0%	1	1	10
q07	22	44	10	10	10	0%	0%	1	1	10
q08	22	44	11	9.97	9	2%	0%	1	0.98	9
q09	22	44	13	9.9	9	14%	0%	1	0.927	9
q10	22	44	13	9.97	8	5%	0%	1	0.872	8

Table 4: AlltheWeb – second period

query	# days monitored	# data comparisons	# URLs identified during period	average overlap	min overlap	% of points overlap less than 10	% of points where relative ranking changed	min Spearman	min G	overlap between first and last point
q01	22	43	20	9.56	6	35%	28%	0.889	0.636	5
q02	22	43	17	9.65	7	30%	12%	0.929	0.836	6
q03	22	43	17	9.65	8	28%	23%	0.842	0.818	7
q04	22	43	28	8.37	5	54%	21%	0.4	0.418	7
q05	22	43	13	9.88	9	12%	26%	0.903	0.909	9
q06	22	43	14	9.77	9	23%	2%	0.933	0.818	8
q07	22	43	15	9.81	8	16%	58%	0.612	0.854	8
q08	22	43	19	9.49	7	35%	23%	0.905	0.745	6

q09	22	43	14	9.77	9	23%	14%	0.85	0.855	9
q10	22	43	20	9.7	2	14%	12%	-1	0.109	1

Table 5: Google.com – second period

Comparing Two Engines

At the time of the data collection the two search engines reportedly indexed approximately the same number of documents (approximately 3 billion documents). In spite of this the results show that the overlap between the top ten results is extremely small (see Tables 6 and 7). The small positive and the negative values of Spearman's rho indicate that the relative rankings on the overlapping elements are considerably different – thus even for those URLs that are considered highly relevant for the given topic by both search engines; the agreement on the relative importance of these documents is rather low.

query	# days monitored	# comparisons	average overlap	min overlap	max overlap	average Spearman	min Spearman	max Spearman	average G	min G	max G
q01	12	21	2	2	2	-1	-1	-1	0.145	0.145	0.145
q02	17	35	4	4	4	0.2978	0.266	0.311	0.4	0.4	0.4
q03	26	110	4.43	4	6	-0.139	-0.8	0.527	0.387	0.245	0.472
q04	24	101	1	1	1	n/a	n/a	n/a	0.177	0.173	0.182
q05	21	42	3	3	3	0.5	0.5	0.5	0.220	0.200	0.267

Table 6: Comparing the search results for AlltheWeb and Google.com – first period

query	# days monitored	# comparisons	average overlap	min overlap	max overlap	average Spearman	min Spearman	max Spearman	average G	min G	max G
q01	22	44	2	2	2	-1	-1	-1	0.133	0.109	0.145
q02	22	44	3.48	2	4	0.361	0.2	1	0.352	0.255	0.418
q03	22	44	3.75	3	5	0.545	0	0.8	0.317	0.291	0.345
q04	22	44	1.05	1	2	n/a	-1	n/a	0.140	0.127	0.236
q05	22	44	1.82	1	2	n/a	n/a	1	0.216	0.182	0.236
q06	22	44	5	5	5	0.698	0.6	0.7	0.6	0.509	0.616
q07	22	44	4.95	4	5	0.202	0.1	0.5	0.416	0.4	0.472
q08	22	44	3.3	2	4	0.493	-1	1	0.309	0.218	0.509
q09	22	44	3.09	3	4	0.527	0.5	0.8	0.438	0.436	0.455
q10	22	44	1.55	1	3	n/a	n/a	0.5	0.109	0.036	0.273

Table 7: Comparing the search results for AlltheWeb and Google.com – second period

There are two possible reasons why a given URL does not appear in the top ten results of a search engine: either it is not indexed by the search engine or the engine ranks it after the first ten results. We checked whether the URLs identified by the two search engines during the second period are indexed by the search engine (we ran this check in February 2004). We defined three cases: the URL was in the top ten list of the engine some time during the period (called "top-ten"), it was not in the top ten, but is indexed by the search engine ("indexed") and is not indexed at all ("not indexed"). The results for queries 1-5 appear in Table 8. The results for these five queries show that both engines index most of the URLs located (between 67.6% and 96.6% of the URLs – top-ten and indexed combined), thus it seems that the ranking algorithms of the two search engines are highly dissimilar.

Query	URLs identified	AlltheWeb			Google.com		
		top-ten	indexed	not indexed	top-ten	indexed	not indexed
q01	28	35.7%	42.9%	21.4%	71.4%	25.0%	3.6%
q02	24	45.8%	45.8%	8.4%	70.8%	25.0%	4.2%
q03	22	50.0%	31.8%	18.2%	77.3%	13.6%	9.1%
q04	39	33.3%	35.9%	30.8%	71.8%	12.8%	15.4%
q05	20	50%	25%	25%	60%	30%	10%

Table 8: URLs indexed by both engines

During the second period we collected data not only from Google.com, but from Google.co.uk and Google.co.il as well, overall the results are rather similar, but there are some differences as can be seen from the results for five randomly chosen queries comparing Google.co.il and AlltheWeb (Table 9 – compare with Table 7) and comparing Google.com with Google.co.il (Table 10).

query	# days monitored	# comparisons	average overlap	min overlap	max overlap	average Spearman	min Spearman	max Spearman	average G	min G	max G
q02	22	44	3.27	3	4	0.42	0.2	0.5	0.37	0.327	0.418
q04	22	44	1.02	1	2	n/a	-1	n/a	0.1	0.127	0.236
q06	22	44	5	5	5	0.602	0.6	0.7	0.6	0.509	0.618
q07	22	44	4.98	4	5	0.237	0.2	0.8	0.406	0.364	0.455
q08	22	44	3.51	3	4	0.749	0.4	1	0.383	0.309	0.436

Table 9: Comparing the search results for AlltheWeb and Google.co.il – second period

query	# days monitored	# comparisons	average overlap	min overlap	max overlap	average Spearman	min Spearman	max Spearman	average G	min G	max G
q01	22	44	9.6	9	10	0.998	0.964	1	0.97	0.909	1
q02	22	44	8.3	3	10	0.987	0.429	1	0.837	0.382	1
q05	22	44	9.75	9	10	0.944	0.745	1	0.95	0.836	1
q06	22	44	9.91	9	10	1	1	1	0.995	0.909	1
q10	22	44	9.98	9	19	0.996	0.903	1	0.996	0.927	1

Table 10: Comparing the search results for Google.com and Google.co.il – second period

Table 10 shows that usually the correlation between google.com and google.co.il is very high – for some reason query 2 (Web data mining) was an exception.

Comparing Two Periods

The second period of data collection took place about three months after the first one. We tried to assess the changes in the top ten lists of the two search engines. The findings are summarized in Table 11. Here we see again that AlltheWeb is less dynamic than Google, except for query 4 (web personalization), where considerable changes were recorded for AlltheWeb as well.

query	AlltheWeb					Google				
	URLs (two periods)	overlap	URLs missing from second set	min change average ranking	max change average ranking	URLs (both period)	overlap	URLs missing from second set	min change average ranking	max change average ranking
q01	11	10	1	0	0.75	22	9	2	0	2.72
q02	11	10	0	0	1	19	10	2	0	5.61
q03	22	8	4	0	2.45	19	12	2	0.18	3.64
q04	19	7	7	0	2.68	32	10	4	0	2.52
q05	10	10	0	0	0	13	10	0	0	1.40

Table 11: Comparing the two periods

Discussion and Conclusions

In this paper, we computed a number of measures in order to assess the changes that occur over time to the rankings of the top ten results on a number of queries for two search engines. We computed a number of measures, since none of them were satisfactory as a standalone measure for such assessment. Overlap does not assess rankings at all, while Spearman's rho ignores the non-overlapping elements and takes into account relative placement only. Moreover, Fagin's measure gives too much weight to the non-overlapping elements. The three measures together provide a better picture than any of these measures alone. Since none of these measures are completely satisfactory, we recommend experimenting with additional measures in the future.

The results indicate that the top ten results usually change gradually. Abrupt changes were observed only very occasionally. Overall, AlltheWeb seems to be much less dynamic than Google. The ranking algorithms of the two search engines seem to be highly dissimilar: even though both engines index most of the URLs that appeared in the top ten lists; the differences in the top ten lists are large (the overlap is small and the correlations between the rankings of the overlapping elements are usually small, sometimes even negative). One reason for Google being more dynamic may be due to its search indexes being unsynchronised while they are being updated, and the non-deterministic nature of query processing due to its distributed nature.

An additional area for further research, along the lines of the research carried out by Vaughan (to appear), is comparing the rankings provided by the search engines with human judgments placed on the value of the retrieved documents.

References

AlltheWeb (2004). Retrieved February 18, 2004 from http://www.alltheweb.com

Chowdhury, A. and Soboroff, I. (2002). Automatic evaluation of World Wide Web Search Services. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, 421-422.

Fagin, R., Kumar, R. and Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1), 134-160.

Google. (2003a*). Google information for Webmasters*. Retrieved February 18, 2004, from http://www.google.com/webmasters/seo.html

Google. (2003b*). Google information for Webmasters*. Retrieved February 18, 2004, from http://www.google.com/webmasters/4.html

Google. (2004) Retrieved February 18, 2004 from http://www.google.com

Hawking, D., Craswell, N., Bailey, P. and Griffiths, K. (2001). Measuring search engine quality. Information Retrieval, 4, 33-59.

Hawking, D., Craswell, N., Thistlewaite, P. and Harman, D. (1999). Results and challenges in Web search evaluation. In *Proceedings of the 8th International World Wide Web Conference*, May 1999, Computer Networks, 31(11-16), 1321-1330, Retrieved February 18, 2004, from http://www8.org/w8-papers/2c-search-discover/results/results.html

Lawrence, S., & Giles, L. (199). Accessibility of information on the Web. *Nature*, 400, 107-109.

Nielsen/NetRatings (2003). *NetView usage metrics*. Retrieved February 18, 2004, from http://www.netratings.com/news.jsp?section=dat_to

Price, G. (2004). Google ups total page count. In *Resourceshelf*. Retrieved February 18, 2004, from http://www.resourceshelf.com/archives/2004_02_01_resourceshelf_archive.html#107702946623981034

Quint, B. (2003). OCLC Project Opens WorldCat Records to Google. In *Information Today*. Retrieved February 18, 2004, from http://www.infotoday.com/newsbreaks/nb031027-2.shtml

Silverstein, C., Henzinger, M., Marais, H and Moricz, M. (1999). Analysis of a very large Web search engine query log. *ACM SIGIR Forum*, 33(1). Retrieved February 18, 2004 from http://www.acm.org/sigir/forum/F99/Silverstein.pdf

Singhal, A., and Kaszkiel, M. (2001). A case study in Web search using TREC algorithms. In *Proceedings of the 10th International World Wide Web Conference*, May 2001, 708-716. Retrieved February 18, 2004 from http://www10.org/cdrom/papers/pdf/p317.pdf

Spink, A., Ozmutlu, S., Ozmutlu, H., C., & Jansen, B. J. (2002). U.S. versus European Web searching trends. *SIGIR Forum*, Fall 2002. Retrieved February 18, 2004 from http://www.acm.org/sigir/forum/F2002/spink.pdf

Soboroff, I., Nicholas, C. and Cahan, P. (2001). Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM SIGIR conference*, 66-72.

Su, L. T., Chen, H.L. and Dong, X. Y. (1998). Evaluation of Web-based search engines from the end-user's perspective: A pilot study. In *Proceedings of the ASIS Annual Meeting*, 35, 348-361.

Sullivan, D. (2003a). *Buying your way in: Search engine advertising chart*. Retrieved February 18, 2004, from http://www.searchenginewatch.com/webmasters/article.php/2167941

Sullivan, D. (2003b). Florida Google dance resources. Retrieved February 18, 2004 from http://www.searchenginewatch.com/searchday/article.php/3285661

Sullivan, D., & Sherman, C. (2004). *4th Annual Search Engine Watch 2003 Awards*. Retrieved February 18, 2004, from http://www.searchenginewatch.com/awards/article.php/3309841

Vaughan, L. (to appear). New measurements for search engine evaluation proposed and tested. To appear in *Information Processing & Management*. doi:10.1016/S0306-4573(03)00043-8

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36, 697-716.