Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Networks 50 (2006) 1448–1463
Computer Networks
www.elsevier.com/locate/comnet

Methods for comparing rankings of search engine results

Judit Bar-Ilan [a,*], Mazlita Mat-Hassan [b], Mark Levene [b]

[a] Department of Information Science, Bar-Ilan University, 52900 Ramat-Gan, Israel
[b] School of Computer Science and Information Systems, Birkbeck, University of London, London, United Kingdom

Available online 20 December 2005

Abstract

In this paper we present a number of measures that compare rankings of search engine results. We apply these measures to five queries that were monitored daily for two periods of 14 or 21 days each. Rankings of the different search engines (Google, Yahoo! and Teoma for text searches and Google, Yahoo! and Picsearch for image searches) are compared on a daily basis, in addition to longitudinal comparisons of the same engine for the same query over time. The results and rankings of the two periods are compared as well.

Keywords: Search engines; Ranking; Similarity measures; Longitudinal analysis

1. Introduction

In merely 15 years the Web has grown to be one of the major information sources. Searching is a major activity on the Web [1,2], and the major search engines are the most frequently used tools for accessing information [3]. Because of the vast amounts of information, the number of results for a large number of queries is in the thousands, and sometimes even in the millions. On the other hand, user studies have shown [4–7] that users browse through the first few results only. Thus results ranking is crucial to the success of a search engine.

In classical IR (information retrieval) systems results ranking was based mainly on term frequency and inverse document frequency (see for example [8, pp. 29–30]). Web search results ranking algorithms take into account additional parameters such as the number of links pointing to the given page [9,10], the anchor text of the links pointing to the page, the placement of the search terms in the document (terms occurring in the title or header may get a higher weight), the distance between the search terms, popularity of the page (in terms of the number of times it is visited), the text appearing in meta-tags [11], subject-specific authority of the page [12,13], recency in search index, and exactness of match [14].

Search engines compete with each other for users, and Web page authors compete for higher rankings with the engines. This is the main reason that search engine companies keep their ranking algorithms secret, as Google states [10]: "Due to the nature of

* Corresponding author. Tel.: +972 3 531 8351; fax: 972 3 535 3937.
E-mail addresses: barilaj@mail.biu.ac.il (J. Bar-Ilan), azy@dcs.bbk.ac.uk (M. Mat-Hassan), mark@dcs.bbk.ac.uk (M. Levene).

1389-1286/$ - see front matter
doi:10.1016/j.comnet.2005.10.020

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 ...''. In addition, search engines continuously fine-tune their algorithms in order to improve the ranking of the results. Moreover, there is a flourishing search engine optimization industry, founded solely in order to design and redesign Web pages so that they obtain high rankings for specific search terms within specific search engines (see for example Search Engine Optimization, Inc., www.seoinc.com).

It is clear from the above discussion that the top-10 results retrieved for a given query have the best chance of being visited by Web users [4–7]. The main motivation for the research we present herein was to examine the differences in the top-10 results and the specific placement of results among different search engines, in addition to examining the changes over time in the top-10 results for a set of queries of the search engines with largest indexes, which at the time of the first data collection were Google, Yahoo! and Teoma (MSN search came out of beta on 1 February 2005 in the midst of the second round of data collection [15]). We also examined results of image searches on Google image search, Yahoo! image search, and on Picsearch (www.picsearch.com). The searches were carried out daily for about 3 weeks in October and November, 2004 and again in January and February, 2005. Five queries (three text queries and two image queries) were monitored. Our aim was to study changes in the rankings over time in the results of the individual engines, and in parallel to study the similarity (or rather non-similarity) between the top-10 results of these tools. In addition, we examined the changes in the results between the two search periods. A larger number of queries over a longer period of time is obviously desirable, but our resources were limited to monitoring the changes occurring in the rankings of only a limited number of queries within the time period of this study.

The goal of the algorithmic ranking functions, is to rank the most ''relevant'' results first, however relevance is a very problematic notion (for extended discussions see [16,17]). We have no clear notion of what is a ''relevant document'' for a given query, and the notion becomes even fuzzier when looking for ''relevant documents'' relating to the user's information seeking objectives. There are several transformations between the user's ''visceral need'' (a fuzzy view of the information problem in the user's mind) and the ''compromised need'' (the way the query is phrased taking into account the limitations of the search tool at hand) [18]. Some researchers (see for example [19]) claim that only the user with the information problem can judge the relevance of the results, while others claim that this approach is impractical (the user cannot judge the relevance of large numbers of documents) and suggest the use of judges or a panel of judges (e.g., in the TREC Conferences, the instructions for the judges appear in [20]). On the Web the question of relevance becomes even more complicated as users usually submit very short queries [4–7]. Consider, for example, the query ''organic food''. What kind of information is the user looking for: an explanation about what organic food is, a list of shops where organic food can be purchased (in which geographic location is the shop?), a site from which he/she can order organic food items, stories about organic food, medical evidence about the advantages of organic food, organic food recipes, and so on. What should the search engine return for such a query and how should it rank the results?

Most previous studies examining ranking of search results base their findings on human judgment. In a study reported by Su [21], users were asked to assess the relevance of the first 20 results retrieved for their queries. In 1999, Hawking et al. [22] evaluated the effectiveness of 20 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 tested and high inter-correlation between the measures. In 2002, Chowdhury and Soboroff [23] also evaluated search effectiveness based on the reciprocal rank; they computed the reciprocal rank of a known item for a query (a URL they a priori paired with the query). In a recent study in 2004, Vaughan [24] 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 and for Teoma only 0.19. Beg [25] compared the rankings of seven search engines on 15 queries with a weighted measure of the users' behavior based on the order the documents were visited,

the time spent viewing them and whether they printed out the document or not. For this study the results of Yahoo!, followed by Google had the best correlation with this measure based on the user's behavior.

Other studies of search results rankings did not involve users. Soboroff et al. [26] based their study on the finding that differences in human judgments of relevance do not affect the relative evaluated performance of the different systems [27]. They proposed a ranking system based on randomly selecting "pseudo-relevant" documents.

Zhao [28] submitted the query "cataloging department" to Google once a week for a period of 10 weeks and studied the changes in the ranks of the 24 sites that were among the top-20 pages during the data collection period. All but three Web sites changed their position at least once during the observation period. The goal of Zhao's study was to try to understand how different parameters (e.g., PageRank, placement of keywords, structure of Website) influence placement, and she provided descriptive statistics to that effect. Eastman and Jansen [29] compared the rankings of search results for queries with and without Boolean operators using the same search terms. In most cases there were no considerable differences between the ranking and coverage (i.e., the number of reported search results). Bifet et al. [30] tried to analyze the factors used in Google's ranking; they found that the parameters influencing the rankings are dependent on the query topic. Joachims [31] claims that under mild assumptions rankings based on clickthrough data give similar results as traditional relevance judgments. Finally, Bar-Ilan recently published a study that compares the rankings of different search engines on the full result sets (i.e., not restricted to the top-10 results) [32]. In that study the comparison was based on Spearman's correlation coefficient, and not on the measures used in the current study.

Fagin et al. [33] introduced a measure (described in the following section) to compare rankings of the top-k results of two search engines, even if the two lists of retrieved documents are not identical. The two lists may contain non-identical documents for two reasons: (1) since only the top-k results are considered, the search engine may have ranked the document after the kth position, and (2) since the search engine has not indexed the given document (It is well-known that the overlap between the indexes of the different search engines is relatively small, see [34–36]. A new study published by Dogpile [37] shows that the overlap on the top-k results is small, and that the top-10 lists of different search engines are extremely different.).

In a previous study [38], we compared the rankings of Google and AlltheWeb on several queries, by computing the size of the overlap, the Spearman correlation on the overlapping elements and a normalized Fagin measure. Each of these measures have their shortcomings (see next section), and thus besides the previous measures, we introduce herein an additional measure for comparing rankings. Two of the queries examined in this paper were also monitored in the previous work.

The aim of the current study is to examine changes in rankings of the top-10 results over time in a given search engine and to compare the rankings provided by different search engines using several comparison measures. The goals of such a study are to gain a better understanding of how different in practice are the ranking algorithms of the different search engines, and to be able to measure the changes in rankings over time.

2. The measures

We used four measures in order to assess the changes over time in the rankings of the search engines and to compare the results of the different search engines. The first and simplest measure is simply the *size of the overlap* between two top-10 lists. Since our other measures are independent of k (top-k lists), we normalize this measure as well, to be the size of the overlap divided by k.

The second measure was Spearman's footrule [39,40]. Spearman's footrule is applied to two rankings of the same set; if the size of the set is N, all the rankings must be between 1 and N (the measure is based on permutations, and thus no ties are allowed). Since the top-10 results retrieved by two search engines for a given query, or retrieved by the same engine on two consecutive days are not necessarily identical, the two lists had to be transformed before Spearman's footrule could be computed. First the non-overlapping URLs were eliminated from both lists, and then the remaining lists were re-ranked; each URL was given its relative rank in the set of remaining URLs in each list. The result of the re-rankings are two permutations σ_1 and σ_2 on $1 \ldots S$, where $|S|$ is the number of overlapping URLs. After these transformations Spearman's footrule is computed as

$$\mathrm{Fr}^{|S|}(\sigma_1,\sigma_2)=\sum_{i=1}^{|S|}|(\sigma_1(i)-\sigma_2(i))|.$$

When the two lists are identical, $\mathrm{Fr}^{|S|}$ is zero, and its maximum value is $1/2|S|^2$ when $|S|$ is even, and $1/2(|S|+1)(|S|-1)$ when $|S|$ is odd. If we divide the result by its maximum value, $\mathrm{Fr}^{|S|}$ will be between 0 and 1, independent of the size of the overlap; we note that this is defined only for $|S|>1$; this measure is undefined for $|S|=0$, 1. Thus we compute the *normalized Spearman's footrule*, NFr, for $|S|>1$

$$\mathrm{NFr}=\frac{\mathrm{Fr}^{(|S|)}}{\max \mathrm{Fr}^{(|S|)}}.$$

NFr ranges between 0 and 1; it attains the value 0 when the two lists are identically ranked and the value 1 when the lists appear in opposite order.

Our other measures are also in this range, but get the value 1 when the lists are identical and the value 0 when they are completely dissimilar. In order to be able to compare the results to those using other measures, we introduce F as

$$F=1-\mathrm{NFr}.$$

Note that Spearman's footrule is based on the re-ranked lists, and thus, for example, if the original ranks of the URLs that appear in both lists (i.e., the overlapping pairs) are (1,8), (2,9) and (3,10), the re-ranked pairs will be (1,1), (2,2) and (3,3) and the value of F will be 1.

The third measure we utilized was one of the metrics introduced by Fagin et al. [33]. 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, Spearman's rho or Spearman's footrule can easily be used. 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. [33] extended the previously mentioned metrics. Here we discuss only the extension of Spearman's footrule, noting that the extension of Kendall's tau was shown in their paper to be equivalent to the extension of Spearman's footrule. We note that 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, given by $\mathrm{Fr}(\sigma_1,\sigma_2)=\sum|\sigma_1(i)-\sigma_2(i)|$. This metric is extended to the case where the two lists are not identical, by assigning an arbitrary placement (which is larger than the length of the list) to documents appearing in one of the lists but not in the other; 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, although 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 now 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, 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. [33] is that when the two lists have little in common, the documents that are not common to the lists have a major effect on the measure. Our experiments show that usually the overlap between the top-10 results of two search engines for an identical query is very small, and thus the non-overlapping elements have a major effect on the measure.

$F^{(k+1)}$ was normalized by Fagin et al. [33] so that the values lie between 0 and 1. For $k=10$ the normalization factor is 110. Thus we compute

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

which we refer to as the *G measure*.

As mentioned above, *Spearman's footrule* is calculated on the re-ranked list of overlapping elements, and ignores the actual rank of the overlapping elements. Thus for the case where there are only two overlapping elements, it cannot differentiate between the cases where the original placements of the overlapping elements are, say

1. (1,1), (2,2),
2. (1,9), (2,10), or
3. (1,2), (2,10).

In all three cases F is 1, since after the re-ranking in all three case we are considering the pairs (1,1) and (2,2). This measure is especially problematic for very small overlaps (say an overlap of size 2), which are typical when comparing the top-10 results

of two search engines on the same query. If the size of the overlap is 2, then there is either total agreement (and F is 1) or total disagreement (and F is 0).

The G measure *does* take into account the placement of the overlapping elements in the lists. For the above examples, the values of G will be:

1. 0.345,
2. 0.055,
3. 0.182.

The G measure seems to capture our intuition that even though the overlapping elements appear in the same order in the two lists, if these appear in places which are more similar, the distance between the measures should be smaller. On the other hand, even if the top five documents are identical, and there is no additional overlap between the lists, the G measure will be 0.727 if the identical elements are in the same order, and 0.618 if they appear in opposite order, i.e., the amount of change in G for a given overlap is rather small and is mainly determined by the size of the overlap.

For this reason we decided to experiment with an additional measure, which we call M. This measure tries to capture the intuition that identical or near identical rankings among the top documents (say the top three documents) is more valuable to the user than among the lower placed documents. In this context Enquiro [41] found in an eye-tracking study, that the top three results were scanned by 100% of the users, while the 10th result was only scanned by 20% of the participants. First, let

$$M' = \sum_{Z} \left| \frac{1}{\mathrm{rank}_1(i)} - \frac{1}{\mathrm{rank}_2(i)} \right| + \sum_{S} \left(\frac{1}{\mathrm{rank}_1(j)} - \frac{1}{(k+1)} \right) + \sum_{T} \left(\frac{1}{\mathrm{rank}_2(j)} - \frac{1}{(k+1)} \right),$$

where Z is the set of the overlapping elements, $\mathrm{rank}_1(i)$ is the rank of document i in the first set and $\mathrm{rank}_2(i)$ is its rank in the second set (both ranks are defined for elements belonging to Z). In addition, S is the set of documents that appear in the first list but not in the second, while T is the set of elements that appear in the second list, but not in the first. These documents may appear in the other list as well, but their rank will be $k+1$ or higher (since we consider the top-10 results only); this is the reason that we subtract $1/(k+1)$ from the reciprocal value of their rank. This measure differs from G, in that it gives a higher weight to higher ranking documents. We have to normalize this measure as well, and to make sure that for identical lists the value of the measure is 1 and for lists, where the overlap is k, and the documents appear in opposite order, the value is 0. The normalizing factor for $k = 10$ is 4.03975. Thus we let

$$M = 1 - \frac{M'}{4.03975}.$$

To demonstrate the difference between the emphases of G and M, assume that the two lists are identical, except that

(a) the first document is different in the two lists. In this case G will be 0.818 and M will be 0.5499;
(b) the last document is different in the two lists. In this case G will be 0.9818 and M will be 0.9955.

Note that in both cases the F-value will be 1. Let us compute the values of M for the examples we used for comparing F with G, i.e.,

1. (1,1), (2,2),
2. (1,9), (2,10), and
3. (1,2), (2,10).

Table 1
Comparing F, G and M

	F	G	M
(1,1), (2,2)	1	0.345	0.653
(1,9), (2,10)	1	0.055	0.015
(1,2), (2,10)	1	0.182	0.207

From Table 1, we can see that in the first case, when the overlapping elements are in high positions in both sets, M is considerably higher than G. On the other hand, when the overlapping elements are in top ranks in the first list but appear at the bottom of the second list, M is much lower than G. Thus we see that M captures our intuition and gives higher weights to higher ranking overlapping elements.

3. Data collection

The data collection for the first round was carried out by six students. The number of queries we were able to monitor was limited by the number of students who carried out this assignment. Their assign-

Table 2
Data collection

Query	First period, # days	Dates first period	Second period, # days	Dates second period
1. US elections 2004	9	1–15 November 2004	21	24 January–13 February, 2005
2. DNA evidence	21	22 October–11 November 2004	21	24 January–13 February
3. Organic food	17	23 October–8 November 2004	21	24 January–13 February
4. Twin towers	24	22 October–15 November 2004	21	24 January–13 February
5. Bondi beach	18	22 October–8 November 2004	21	24 January–13 February

ment involved choosing a text query and an image query from a given list of queries and to submit these queries to the appropriate search engine once a day for a period of 14 days. The students started data collection at different dates; therefore if two or more students monitored the same query, we had data for these queries for more than 14 days. Sometimes the students skipped a day of data collection, or there was no overlap between the students' work, and thus the data for the first period is not completely continuous. During the second period all the queries were monitored for 21 consecutive days by one of the authors. Table 2 displays the queries, the number of days these queries were monitored and the time span of the data collection for each period.

The choice of queries that were submitted to the search engines were decided a priori by the authors. For the text queries, we chose a query to represent a topical topic (US elections 2004) and two queries from our previous study [38]; organic food and DNA evidence. These two queries were chosen as we were interested in investigating whether previously monitored URLs were still available during the current observation period. For image queries, the queries were chosen to represent *places* (Bondi beach and Twin towers) and an *event* (Twin towers). The query "Twin towers" was particularly interesting as it represents both a place and an event (the 9/11 attacks on the World Trade Centre).

The first three queries were text searches and were submitted at each data collection point to Google, Yahoo! and Teoma (the largest search engines in terms of their index at the time of data collection), while the last two queries were image searches and were submitted to Google image search, Yahoo! image search and to Picsearch (www.picsearch.com). Google image and Yahoo! image were the largest image search engines in terms of size at the time of data collection, the choice of the third engine, Picsearch was made because it was delivering image searches to MSN Search (beta) [42] and Ask Jeeves at the time of data collection [43]. The URLs and the rankings of the top-10 results for each query and for each search engine were recorded at each data collection point. For the image searches, the URLs of the images (and not of the embedding pages) were recorded.

4. Data analysis

For a given search engine and a given query we computed the overlap (O), Spearman's footrule (F), Fagin's measure (G) and our new measure (M), on the results for consecutive data collection points. The results of pairs of engines were also compared by computing the same measures for the two ranked lists retrieved by the two search engines on the same day, for each day recorded. The two periods were compared on all five queries; we calculated the overlap between the two periods and assessed the changes in the rankings of the overlapping elements based on the average rankings. For all the queries, the maximum values of all the measures were 1, except for "US elections 2004", where the maximum overlap was only 0.9 for Google, noting that, at times, the data was not collected on consecutive days. An additional reason for this could be that the data was created very close to the elections (between 1 and 15 November 2004; the elections were held on 2 November 2004).

5. Results

5.1. The first round

As can be seen from Table 3, Google's set of results and rankings fluctuated slightly during the period of data collection, with the exception of the query "organic food" which was very stable; the same is true of the other two search engines for this query. Even though Google covered 13 URLs among the top-10 results for the query "Bondi beach", the result sets for the first and last day were identical.

Table 3
Measures for the changes in ranking of the individual engines over time—round 1

Query + search engine	Overlap		*F*		*G*		*M*		#URLs	Overlap between first and last day
	Avg	Min	Avg	Min	Avg	Min	Avg	Min		
US elections 2004										
Google	0.81	0.7	0.86	0.51	0.78	0.6	0.74	0.74	18	8
Yahoo!	0.93	0.8	0.95	0.84	0.92	0.8	0.95	0.86	15	8
Teoma	1	1	1	1	1	1	1	1	10	10
Organic food										
Google	1	1	0.95	0.80	0.98	0.91	0.99	0.95	10	10
Yahoo!	1	1	1	1	1	1	1	1	10	10
Teoma	0.99	0.9	1	1	0.99	0.95	0.99	0.98	11	9
DNA evidence										
Google	0.91	0.8	0.98	0.88	0.93	0.84	0.97	0.91	18	7
Yahoo!	1	1	1	1	1	1	1	1	10	10
Teoma	0.96	0.9	0.95	0.80	0.94	0.85	0.96	0.84	12	9
Twin towers										
Google	0.93	0.7	0.88	0.5	0.89	0.62	0.92	0.70	13	7
Yahoo!	0.96	0.8	0.94	0.56	0.95	0.78	0.95	0.75	14	10
Picsearch	0.98	0.5	1	1	0.98	0.67	0.99	0.85	14	5
Bondi beach										
Google	0.81	0.7	0.81	0.43	0.84	0.62	0.89	0.74	13	10
Yahoo!	0.88	0.2	0.92	0	0.86	0.15	0.84	0.05	21	2
Picsearch	0.99	0.9	0.99	0.96	0.98	0.82	0.98	0.78	11	9

Fig. 1 depicts the changes in the placements and occurrence of the URLs during the data collection period for the query "DNA evidence". We see from the figure, that the top-three places were stable during the whole period. URL4 (http://books.nap.edu/html/DNA) was ranked fourth for 8 days, then appeared in the top-10 for five additional days (in places five and six) and then disappeared from the top-10 (although it continued to exist, and as we shall see later, it reappeared in the top-10 during the second data collection period). The fourth place was taken by URL5 (www.nap.edu/catalog/5141.html), which was initially ranked number 5. Both pages contain information on a 1996 publica-

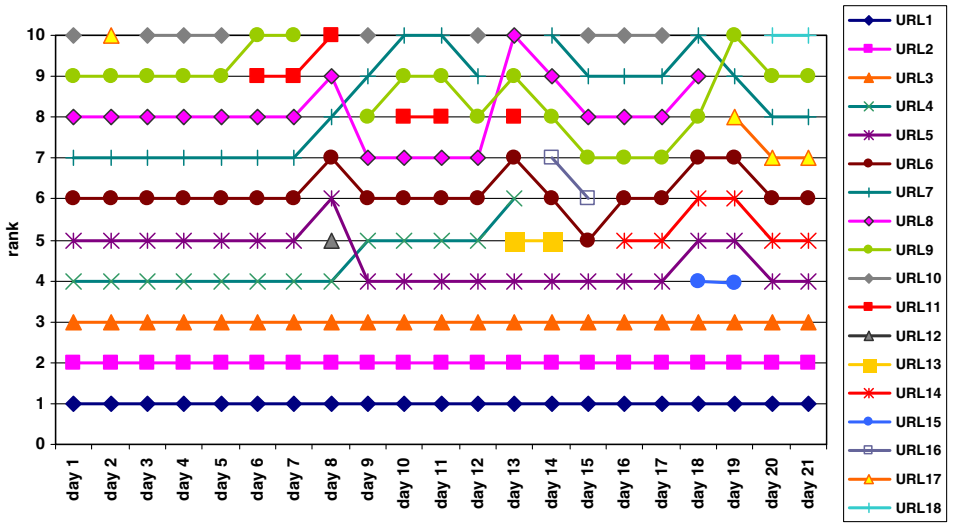

Fig. 1. The top-10 results of Google for the query "DNA evidence".

tion of the Committee of DNA Forensic Science of the US National Research Council, although URL4 has much more actual content, while URL5 offers the purchase of the report. URL1 (www.howstuffworks.com/dna-evidence.htm) presents popular information, and is identical in content to URL15, which appeared as number 4 on days 19 and 20. URLs 2 and 3 (www.ojp.usdoj.gov/nij/dna and www.ojp.usdoj.gov/nij/dna_evbro) contain information provided by the US Department of Justice on the topic.

Yahoo! showed either minor or no changes in the results of the text queries, and was also relatively stable for the image query "Twin towers", however its results fluctuated heavily for the query "Bondi beach"—only two URLs appeared among the top-10 for all 18 days. We could identify four different sub-periods, as can be seen in Table 4.

Teoma was highly stable during the period. Picsearch retrieved almost identical results on "Bondi beach" at each data collection point, while for the query "Twin towers" there was a single, but considerable change on 10 November 2004, otherwise the results and rankings were stable.

We also compared the rankings of the different search engines on the same query on the same day, using the same measures. There was no overlap between the top-10 results of any of the pairs of the search engines for the image query "Bondi beach". The situation was almost the same for the image query "Twin towers", except for a single image that appeared as #1 on Google during the whole period, and fluctuated between places 7 and 9 on Yahoo!. Therefore, Table 5 displays the measures (average, minimum and maximum values) for all the text queries and for the image query "Twin towers" for the pair Google-Yahoo! only.

The highest overlap (seven overlapping URLs in the top-10) was between Google and Yahoo! for the query "organic food". The similarity measures are not very high, because the relative ranking of these URLs by the two engines are somewhat different: #1 on Google is #2 on Yahoo! and visa versa, #4 on Google is #8 on Yahoo!, and #4 on Yahoo! is between the 7th and 10th places on Google. Even though there are only two overlapping elements in the top-10 for the query "organic food" between Yahoo! and Teoma, their rankings are the same (#2 and #3 on both engines), thus the F value is one. The G and M values are relatively low because the size of the overlap is small, M is slightly higher than G, because the ranks of the overlapping elements are relatively high. These two URLs also appear on Google's lists: #2 on Teoma and Yahoo! is #1 on Google, and #3 on Teoma and Yahoo! ranks between 5 and 8 on Google.

There are two cases where there was only a single overlapping URL in the top-10 results

Table 4
Yahoo!'s top-10 results for the image query "Bondi beach"

	Days 1–3	Day 4	Days 5–11	Days 12–18
URL1	1		1	
URL2	2		2	
URL3	3	5		
URL4	4	7		
URL5	5	6	5	6
URL6	6	8	3	8
URL7	7		4	
URL8	8		10 (was #9 on day 7)	
URL9	9	10	7	
URL10	10		6	
URL11		1		3
URL12		2		1
URL13		3		2
URL14		4		4
URL15		9	8	
URL16			9 (was not among the top-10 on day 7)	
URL17			#10 on day 7	
URL18				5
URL19				7
URL20				9
URL21				10

Table 5
Measures for comparing the rankings of the different search engines on identical queries at the same data collection points—first round

Query + search engine	Overlap			*F*			*G*			*M*		
	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max
US elections 2004												
Google–Yahoo!	0.34	0.3	0.4	0.43	0.11	0.56	0.36	0.29	0.44	0.29	0.25	0.35
Yahoo!–Teoma	0.28	0.2	0.3	1	1	1	0.28	0.2	0.33	0.25	0.21	0.27
Google–Teoma	0.48	0.4	0.6	0.63	0.44	0.75	0.42	0.36	0.47	0.36	0.34	0.37
Organic food												
Google–Yahoo!	0.7	0.7	0.7	0.52	0.5	0.58	0.61	0.56	0.69	0.51	0.48	0.55
Yahoo!–Teoma	0.2	0.2	0.2	1	1	1	0.31	0.31	0.31	0.32	0.32	0.32
Google–Teoma	0.3	0.3	0.3	0.88	0.5	1	0.28	0.24	0.31	0.26	0.22	0.29
DNA evidence												
Google–Yahoo!	0.39	0.3	0.4	1	1	1	0.50	0.45	0.53	0.66	0.65	0.67
Yahoo!–Teoma	0.3	0.3	0.3	0.5	0.5	0.5	0.33	0.31	0.35	0.51	0.50	0.52
Google–Teoma	0.1	0.1	0.1	N/A	N/A	N/A	0.18	0.18	0.18	0.45	0.45	0.45
Twin towers												
Google–Yahoo!	0.1	0.1	0.1	N/A	N/A	N/A	0.05	0.04	0.09	0.02	0.01	0.04

(Yahoo!–Teoma for "DNA evidence", and Google–Yahoo! for "Twin towers"): the considerable difference between the *G* values for these two cases are caused because of the different ranks of the overlapping element. For "DNA evidence" both engines ranked the overlapping URL as #1, while for "Twin towers", Google ranked the overlapping URL as #1, and Teoma's rank for this URL varies between 6 and 9.

Let us take a closer look at the query "DNA evidence". All three engines agree on the top-ranked URL (www.howstuffworks.com/dna-evidence.htm) for the whole period. Google and Yahoo! overlap on four URLs, except for a single day where there were only three overlapping URLs. These URLs were constant during the whole period: the first three are URLs #1–#3 of Google (mentioned above)—they were ranked 1, 3 and 4 on Yahoo! respectively. We see that there is a high degree of agreement between the two search engines regarding the top results. The fourth overlapping URL is #7 on Yahoo! and fluctuates between ranks 7 and 10 on Yahoo!. Rank 2 on Yahoo! (www.ncjrs.org/txtfiles/dnaevid.txt) is ranked as #10 on Teoma (and is not among the top-10 in Google). Teoma overlaps with Yahoo! on URLs ranked 1, 2 and 5 on Yahoo!'s lists. These URLs are ranked 1, 10 and 4–5 on Teoma.

5.2. The second round

Table 6 summarizes the findings related to the rankings of each search engine over time. For each measure we provided the average and the minimum (over 21 days). The maximum value attained for all four measures was 1.

Unlike in the first round, this time the results for the query "US elections 2004" were rather stable for all the engines. This finding is not surprising, since the second round took place almost 3 months after the elections.

Teoma retrieved exactly the same results in the same order on all 21 days for the query "organic food". On the other hand, Yahoo! image searches had the most fluctuations. Fig. 2 depicts the fluctuations in 10 out of the 20 URLs that were identified by Yahoo! for the query "Bondi beach".

Next we compared the rankings of the different search engines on the same query on the same day. This time, there was no overlap at all between the search engine results for the image queries. Therefore, Table 7 displays the measures (average, minimum and maximum values) for the text queries only.

Let us examine two cases more closely, Google versus Teoma on "organic food", where the *M* values are lower than the *G* values; and Yahoo! versus Teoma on "DNA evidence" where the *M* values are higher than the *G* values.

For the query "organic food" the number of overlapping URLs for Google and Teoma varies between 2 and 4. Two URLs overlap on all days, except one: www.organicfood.co.uk, which is #1 on Google, and #4 on Teoma (and #2 on Yahoo!) and www.rain.org/~sals/my.html, which is between #6 and #9 on Google (and not among the top-10 on day 2) and #3 on Teoma (its rank varied between 3 and 4

Table 6
Measures for the changes in ranking of the individual engines over time—round 2

Query + search engine	Overlap		*F*		*G*		*M*		#URLs located (whole period)	Overlap between first and last day
	Avg	Min	Avg	Min	Avg	Min	Avg	Min		
US elections 2004										
Google	0.97	0.8	0.9	0.3	0.94	0.65	0.88	0.31	12	9
Yahoo!	0.99	0.8	1	1	0.98	0.82	0.99	0.79	13	8
Teoma	0.98	0.8	0.98	0.85	0.97	0.8	0.97	0.8	13	9
Organic food										
Google	0.91	0.5	0.93	0.67	0.92	0.46	0.95	0.57	15	8
Yahoo!	0.99	0.9	1	1	0.99	0.91	0.99	0.96	11	10
Teoma	1	1	1	1	1	1	1	1	10	10
DNA evidence										
Google	0.91	0.8	0.99	0.9	0.94	0.84	0.98	0.9	19	7
Yahoo!	0.99	0.8	1	1	0.99	0.85	0.99	0.92	12	9
Teoma	0.97	0.8	0.98	0.88	0.97	0.82	0.98	0.87	14	8
Twin towers										
Google	0.97	0.6	0.96	0.44	0.96	0.6	0.96	0.45	15	6
Yahoo!	0.88	0.5	0.9	0.3	0.83	0.42	0.74	0.15	20	7
Picsearch	0.98	0.6	1	1	0.98	0.55	0.99	0.72	14	6
Bondi beach										
Google	0.96	0.7	0.99	0.75	0.95	0.69	0.97	0.82	15	9
Yahoo!	0.87	0.5	0.91	0	0.85	0.31	0.81	0.14	20	6
Picsearch	0.99	0.9	0.99	0.96	0.99	0.91	0.99	0.96	11	9

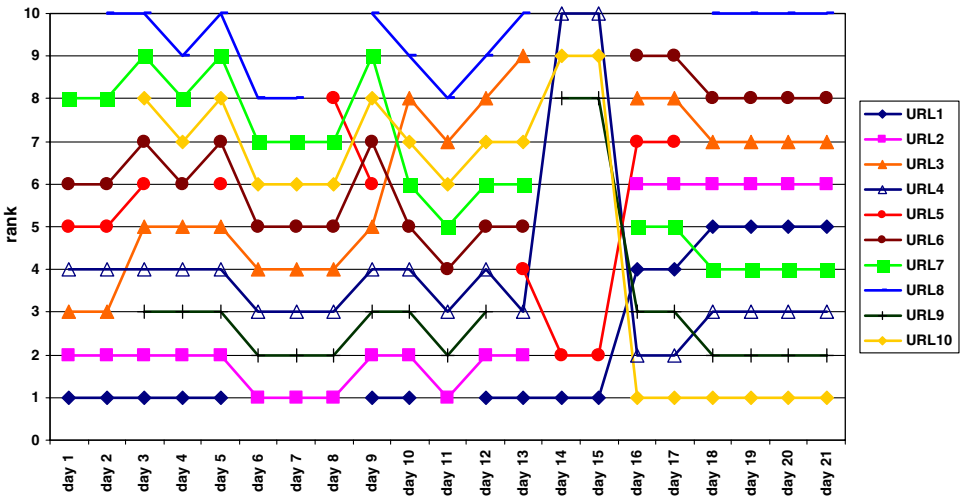

Fig. 2. Fluctuations in rankings of Yahoo! for the query "Bondi beach".

on Yahoo! as well). Thus the *F* values, based on relative rankings only, are low, the *M* value is very low as well, because the number of overlapping URLs is small, and there is considerable disagreement between the rankings, while the *G* value is higher, because it puts more weight on the number of overlapping elements, and less on their relative placements.

We identified five overlapping URLs for Yahoo! and Teoma for the query "DNA evidence". The overlapping URLs are ranked 1, 2, 3, 7 and 8 by Yahoo! and 1, 5–7, 4–5, 2 and 8–10 by Teoma, respectively. The *M* value is relatively high, because the top elements overlap, even though the relative rankings are not the same for the two engines.

Table 7
Measures for comparing the rankings of the different search engines on identical queries at the same data collection point

Query + search engine	Overlap			*F*			*G*			*M*		
	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max
US elections 2004												
Google–Yahoo!	0.4	0.3	0.5	0.19	0	0.5	0.46	0.29	0.55	0.33	0.14	0.47
Yahoo!–Teoma	0.5	0.4	0.6	0.34	0.11	0.75	0.44	0.31	0.49	0.31	0.14	0.38
Google–Teoma	0.52	0.4	0.6	0.58	0.5	0.78	0.46	0.4	0.58	0.35	0.27	0.45
Organic food												
Google–Yahoo!	0.64	0.5	0.7	0.44	0.42	0.67	0.53	0.45	0.55	0.46	0.33	0.48
Yahoo!–Teoma	0.2	0.2	0.2	0	0	0	0.26	0.25	0.27	0.18	0.16	0.2
Google–Teoma	0.31	0.2	0.4	0.36	0	0.75	0.22	0.18	0.31	0.11	0.09	0.14
DNA evidence												
Google–Yahoo!	0.4	0.4	0.4	0.75	0.75	0.75	0.45	0.42	0.46	0.61	0.6	0.62
Yahoo!–Teoma	0.5	0.5	0.5	0.67	0.67	0.67	0.53	0.45	0.55	0.61	0.56	0.63
Google–Teoma	0.2	0.2	0.2	1	1	1	0.38	0.29	0.39	0.52	0.5	0.53

5.3. Comparison between the rounds

The results for the first round were collected from the end of October until the beginning of November 2004. The second round took place 3 months later from the end of January until the beginning of February 2005. In this section we examine how the top-10 results changed during the 3 months gap. The aggregated results are displayed in Table 8.

The average rank of a URL for a search engine in a search round is the sum of the rankings it received on each day the URL appeared among the top-10 results of the search engine for the query, divided by the number of days it appeared in the top-10 list. Thus the average rank of a URL is between 1 and 10. The change in the average rank of a URL is defined as the absolute value of the difference between its average rank in round 1 and round 2

Table 8
Changes to the top-10 results between the rounds

	Query	US elections 2004	Organic food	DNA evidence	Twin towers	Bondi beach
Google	# URLs identified during both periods	19	15	26	20	17
	Overlap	11	10	13	8	11
	# URLs in first set, but missing from second set	7	0	5	4	2
	Min change in average ranking	0.24	0	0	0.34	0
	Max change in average ranking	2.33	7.29	4	4.04	2.77
Yahoo!	# URLs identified during both periods	18	13	12	24	34
	Overlap	9	8	10	8	8
	# URLs in first set, but missing from second set	5	2	0	4	13
	Min change in average ranking	0	0	0	1.53	1.72
	Max change in average ranking	3.63	2.1	5	5.16	4.73
Teoma	# URLs identified during both periods	15	12	21		
	Overlap	8	9	5		
	# URLs in first set, but missing from second set	2	2	7		
	Min change in average ranking	0	0	0		
	Max change in average ranking	4.38	2	3.76		
Picsearch	# URLs identified during both periods				25	18
	Overlap				3	4
	# URLs in first set, but missing from second set				11	7
	Min change in average ranking				3.14	0.63
	Max change in average ranking				6	5.78

(the value is undefined if it was missing from either round). The minimum and maximum values were computed for each search engine and each query over all the URLs for which the change was defined. The smaller the maximum change, the more similar are the rankings in the two rounds.

Here we see again, that the results of Teoma and Yahoo! (text searches) were most stable (the number of URLs identified in the first round, but missing in the second round from the top-10, was the smallest). The query "organic food" had more stable results than the other queries we examined (the least number of URLs identified, and the least number of URLs missing from the second set).

At this point in time, image searches are rather different from text searches. Even though the results of Picsearch were very stable during the each data collection period, the results changed considerably between periods. Google was most stable for image searches. Google admitted in November 2004 that it had not updated its image database for some time [44]. In spite of this report, we still observed considerably changes during the first round in the top-10 results of Google for our queries. On 8 February 2004, Google announced that it had refreshed and expanded its image database [45]. This was in the middle of the second round of data collection. We saw some changes in the top-10 results for Google on 9 February 2005. The overlap with the results of the previous day was only six URLs for "Twin towers", and their relative rankings also changed considerably. On the other hand, only a minor change was observed for "Bondi beach", and we saw more considerable changes in the daily results in the top-10 list before the date of the expansion.

For the text queries, in eight out of nine cases (three search engines, three queries each) there was at least one URL whose average rank had not changed between the search rounds (as can be seen from the rows for minimum change in average ranking). In all of these cases, this minimum was achieved by the top ranking URL, i.e., the top-ranking URL was ranked #1 at each data collection point both in round one and in round two. For the image searches, only for Google, for the query "Bondi beach", did the #1 URL remain the same during both periods.

6. Discussion

The queries "DNA evidence" and "organic food" were also monitored in our previous study [32]. Then we submitted the queries to Google and to AllTheWeb, during two data collection periods: in October 2003 and in January 2004 (i.e., exactly a year before the current data collection rounds). We identified 4 URLs for the query "DNA evidence" and 6 URLs for "organic food" that appeared in all four data collection rounds. Figs. 3 and 4 depict the average rankings of these URLs during the four data collection periods. The rankings for "DNA evidence" are much more stable than the rankings for "organic food". It is

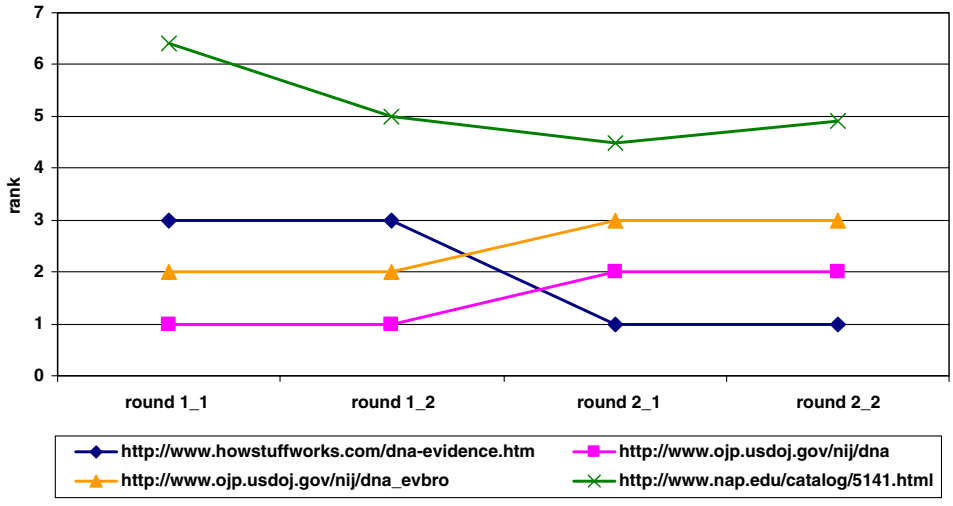

Fig. 3. The changes to the average rankings assigned by Google in the four data collection periods for the query "DNA evidence".

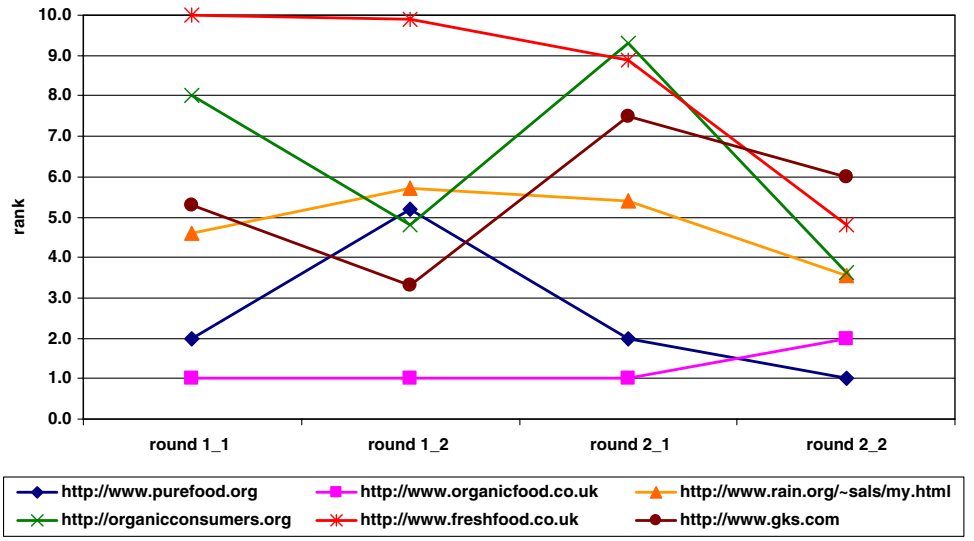

Fig. 4. The changes to the average rankings assigned by Google in the four data collection periods for the query "organic food".

interesting to see that so many URLs remained in the top-10 results for these queries for over a year.

Jux2 was a tool for visualizing the overlap between the top results of Google, Yahoo! and AskJeeves. It reported that on the 500 most popular search terms, the average overlap between Google and Yahoo! was 3.8, and for 30% of the queries the overlap was between 0 and 2 [46]. Jux2 also reported that the overlap between Google and AskJeeves (powered by Teoma) is even smaller, 3.4 on average, while the average overlap between Yahoo! and AskJeeves is only 3.1 on average. The Dogpile study [Do] discusses the overlap between Google, Yahoo!, MSN and AskJeeves on the top-10 results. Their findings show that only 1.1% of the results are shared by all four search engines, 2.6% by three, 11.4% by two, and 84.9% of the top-10 hits were retrieved by a single engine only. For "our" three test queries, the average overlap was 4.8 between Google and Yahoo!, 3.4 between Google and Teoma, and 4 between Yahoo! and Teoma. Thus our results differ from the statistics provided by Jux2 and Dogpile. They tested a much larger set of queries, and the queries observed by us may not have been among the ones they tested.

In this study we also experimented with image queries, an extension to our previous study where only text queries were observed. The queries were specifically chosen to represent a *place* (Bondi beach, Twin towers) and an *event* (Twin towers). Our experiments have shown that while results were very stable during each specific data collection for all search engines involved, the results changed considerably between periods. Overlapping between search engines was either non-existent or minimal. On 17 April 2005 we queried the search engines Google and Yahoo! for the existence of URLs identified by Yahoo! and Google, respectively, during the second search round for the queries "DNA evidence" and "Bondi beach" (for the image query we submitted the URLs in which the images were embedded). Google indexed all 12 URLs identified by Yahoo! for "DNA evidence", but only 7 out of the 20 URLs located by Yahoo! with images on "Bondi beach". Yahoo! covered Google's image searches much better; it indexed 11 out of the 15 URLs located by Google, but did worse on the text query, where it indexed 14 out of the 19 URLs identified during the period by Google on the query "DNA evidence". Note, that we checked the overlap after the latest announced update of Google Images [45]; this update took place in the middle of the second data collection round, thus the URLs located in the second round were partially collected from the new database.

7. Conclusions

We have experimented with a number of measures in order to assess the changes that occur over time to the rankings of the top-10 results of search engines, and to assess the differences in the rankings

of different search engines. In our previous study, we computed the overlap, Spearman's rho and Fagin's G measure. We observed that these measures are not fully satisfactory on their own, and thus we recommended that all of the three measures should be used.

In the current study we computed four measures: the overlap, Spearman's footrule, F, Fagin's G measure, and the new M measure. Our reason for introducing this new measure was to minimize the problems related to the other measures. The overlap ignores rankings, Spearman's footrule is based only on the relative rankings and ignores the non-overlapping elements completely, and, finally, Fagin's measure gives far too much weight to the size of the overlap. The new measure attempts to take into account both the overlapping and the non-overlapping elements, and gives higher weight to the overlapping URLs among the top-ranking results. It seems that the M measure better captures our intuition regarding the quality of rankings, but further studies are needed to show the full utility of this measure (and/or experimenting with additional measures). The recent eye-tracking study [41] supports our intuition that higher weight should be given to overlap in the top results.

We experimented both with text and image queries. Results of image queries were less stable, and the overlap between the results of the different image search tools was non-existent or minimal. This is striking compared to the average overlap of 0.41 between all pairs of search engines for all the text queries. Thus it seems that either there is much more agreement on the "importance" of textual data versus image data, or that the image databases of the different search engines are almost disjoint.

Our results seem to indicate that even though the overlap between the top-ranked documents for the image queries is lower than for text queries, the overlap is still considerable. Thus it seems that the differences in the coverage of the image databases only provide a partial explanation for the different results obtained by the different search tools for the image queries. Further studies are needed in this area as well.

References

[1] M. Madden, America's online pursuits: The changing picture of who's online and what they do, PEW Internet and American Life Project, 2003, Available from: <http://www.pewinternet.org/pdfs/PIP_Online_Pursuits_Final.PDF>.

[2] D. Fallows, The Internet and daily life, PEW Internet and American Life Project, 2004, Available from: <http://www.pewinternet.org/pdfs/PIP_Internet_and_Daily_Life.pdf>.

[3] D. Sullivan, Nielsen Netratings search engine ratings, Search Engine Watch Reports, 2005, Available from: <http://searchenginewatch.com/reports/article.php/2156451>.

[4] C. Silverstein, M. Henzinger, H. Marais, M. Moricz, Analysis of a very large Web search engine query log, ACM SIGIR Forum 33 (1) (1999), Available from: <http://www.acm.org/sigir/forum/F99/Silverstein.pdf>.

[5] A. Spink, S. Ozmutlu, H.C. Ozmutlu, B.J. Jansen, US versus European Web searching trends, SIGIR Forum, 2002, Available from: <http://www.acm.org/sigir/forum/F2002/spink.pdf>.

[6] B.J. Jansen, A. Spink, An analysis of Web searching by European Alltheweb.com users, Information Processing and Management 41 (6) (2004) 361–381.

[7] B.J. Jansen, A. Spink, J. Pedersen, A temporal comparison of AltaVista Web searching, Journal of the American Society for Information Science and Technology 56 (6) (2005) 559–570.

[8] R.A. Baeza-Yates, B.A. Ribeiro-Neto, Modern information retrieval, ACM Press, Addison-Wesley, Harlow, England, 1999.

[9] S. Brin, L. Page, The anatomy of a large-scale hypertextual Web search engine, in: Proceedings of the 7th International World Wide Web Conference, April 1998, Computer Networks and ISDN Systems, vol. 30, 1998, pp. 107–117, Available from: <http://www-db.stanford.edu/pub/papers/google.pdf>.

[10] Google, Google information for Webmasters, 2004, Available from: <http://www.google.com/webmasters/4.html>.

[11] Yahoo!, Yahoo!! Help: How do I improve the ranking of my website in the search results, 2005, Available from: <http://help.yahoo.com/help/us/ysearch/ranking/ranking-02.html>.

[12] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (5) (1999) 604–632.

[13] Teoma (2005), Adding a new dimension to search: The Teoma difference is authority, Retrieved 26 March 2005, Available from: <http://sp.teoma.com/docs/teoma/about/searchwithauthority.html>.

[14] MSN Search, Web search help: change your search results by using results ranking, 2005, Available from: <http://search.msn.com/docs/help.aspx?t=SEARCH_PROC_BuildCustomizedSearch.htm>.

[15] C. Payne, MSN Search launches, 2005, Available from: <http://blogs.msdn.com/msnsearch/archive/2005/01/31/364278.aspx>.

[16] T. Saracevic, RELEVANCE: a review of and a framework for the thinking on the notion in information science, Journal of the American Society for Information Science (1975) 321–343.

[17] S. Mizzaro, Relevance: the whole history, Journal of the American Society for Information Science 48 (9) (1997) 810–832.

[18] R.S. Taylor, Question-negotiation and information seeking in libraries, College and Research Libraries 29 (May) (1968) 178–194.

[19] M. Gordon, P. Pathak, Finding information of the World Wide Web: the retrieval effectiveness of search engines,

Information Processing and Management 35 (1999) 141–180.
[20] TREC, Data—English relevance judgements, 2004, Available from: <http://trec.nist.gov/data/reljudge_eng.html>.
[21] L.T. Su, A comprehensive and systematic model of user evaluation of Web search engines: II. An evaluation by undergraduates, Journal of the American Society for Information and Technology 54 (2003) 1193–1223.
[22] D. Hawking, N. Craswell, P. Bailey, K. Griffiths, Measuring search engine quality, Information Retrieval 4 (2001) 33–59.
[23] A. Chowdhury, I. Soboroff, Automatic evaluation of World Wide Web Search Services, in: Proceedings of the 25th Annual International ACM SIGIR Conference, 2002, pp. 421–422.
[24] L. Vaughan, New measurements for search engine evaluation proposed and tested, Information Processing and Management 40 (2004) 677–691.
[25] M.M.S. Beg, A subjective measure of Web search quality, Information Sciences 169 (2005) 365–381.
[26] I. Soboroff, C. Nicholas, P. Cahan, Ranking retrieval systems without relevance judgments, in: Proceedings of the 24th Annual International ACM SIGIR Conference, 2001, pp. 66–72.
[27] E.M. Voorhees, Variations in relevance judgments and the measurement of retrieval effectiveness, Information Processing and Management 36 (2000) 697–716.
[28] L. Zhao, Jump higher: analyzing Web-site rank in Google, Information Technology and Libraries 23 (4) (2004) 108–118.
[29] C. Eastman, B.J. Jansen, Coverage, relevance and ranking: the impact of query operators on Web search engine results, ACM Transactions on Information Systems 24 (2003) 383–411.
[30] A. Bifet, C. Castillo, P.A. Chirita, I. Weber, An analysis of factors used in search engine ranking, AIRWeb'05, Available from: <http://airweb.cse.lehigh.edu/2005/bifet.pdf>.
[31] T. Joachims, Evaluating retrieval performance using clickthrough data, in: Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval, 2002, Available from: <http://www.cs.cornell.edu/People/tj/publications/joachims_02.pdf>.
[32] J. Bar-Ilan, Comparing rankings of search results on the Web, Information Processing and Management 41 (2005) 1511–1519.
[33] R. Fagin, R. Kumar, D. Sivakumar, Comparing top k lists, SIAM Journal on Discrete Mathematics 17 (1) (2003) 134–160.
[34] K. Bharat, A.Z. Broder, A technique for measuring the relative size and overlap of public Web search engines, in: Proceedings of the 7th International World Wide Web Conference, April 1998, Computer Networks and ISDN Systems, vol. 30, 1998, pp. 379–388, Available from: <http://www.ra.ethz.ch/CDstore/www7/1937/com1937.htm>.
[35] S. Lawrence, C.L. Giles, Accessibility of information on the Web, Nature 400 (1999) 107–109.
[36] A. Gulli, A. Signorini, The indexable Web is more than 11.5 billion pages, in: Proceedings of the WWW2005 Conference, May 2005, Available from: <http://www2005.org/cdrom/docs/p902.pdf>.
[37] Dogpile, Different engines, different results, Available from: <http://compareseachengines.dogpile.com/OverlapAnalysis.pdf>.
[38] J. Bar-Ilan, M. Levene, M. Mat-Hassan, Dynamics of search engine rankings—A case study, in: Proceedings of the 3rd International Workshop on Web Dynamics, New York, May 2004, Available from: <http://www.dcs.bbk.ac.uk/webDyn3/webdyn3_proceedings.pdf>.
[39] P. Diaconis, R.L. Graham, Spearman's footrule as a measure of disarray, Journal of the Royal Statistical Society, Series B (Methodological) 39 (1977) 262–268.
[40] C. Dwork, R. Kumar, M. Naor, D. Sivakumar, Rank aggregation methods for the Web, in: Proceedings of the 10th World Wide Web Conference, Hong-Kong, May 2001, pp. 613–622.
[41] Enquiro, Did-It, Enquiro and Eyetools uncover search's golden triangle, Available from: <http://www.enquiro.com/eye-tracking-pr.asp>.
[42] C. Sherman, Microsoft unveils its new search engine at last, Searchday, 11 November 2004, Available from: <http://searchenginewatch.com/searchday/article.php/3434261>.
[43] C. Sherman, Ask Jeeves serves up new features, Searchday, 21 April 2003, Available from: <http://searchenginewatch.com/searchday/article.php/2194051>.
[44] D. Sullivan, Stale and split image databases fuel Google conspiracy, November 2004, Available from: <http://blog.searchenginewatch.com/blog/041108-145734>.
[45] D. Sullivan, Google Images updates, expands, enters main results, February 2005, Available from: <http://blog.searchenginewatch.com/blog/050207-155007>.
[46] S. Spencer, The overlap between Google and Yahoo! Results is less than you might think, Natural Search Blog, 29 August 2004, Available from: <http://www.naturalsearchblog.com/archives/2004/08/29>.

Judit Bar-Ilan is a senior lecturer at the Department of Information Science of the Bar-Ilan University, Israel. She received her Ph.D. in Computer Science from the Hebrew University of Jerusalem. She started her research in information science in the mid-1990s. Her areas of interest include: information retrieval, informetrics, the semantic Web, Internet research, information behavior and usability.

Mazlita Mat-Hassan is a Ph.D. student at the School of Computer Science and Information Systems of Birkbeck, University of London, United Kingdom. Her areas of interest include: Web data mining, Web information retrieval, user modeling, usability and search and navigation behaviour.

Mark Levene received his Ph.D. in Computer Science in 1990 from Birkbeck College, University of London, having previously been awarded a BSc in Computer Science from Auckland University, New Zealand in 1982. He is currently Professor of Computer Science at Birkbeck College, where he is a member of the Information Management and Web Technologies research group. His main research interests are Web search and navigation, Web data mining and stochastic models for the evolution of the Web. He has published extensively in the areas of database theory and Web technologies, and has recently published a book called An Introduction to Search Engines and Web Navigation.