Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Processing and Management 41 (2005) 1511–1519

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Comparing rankings of search results on the Web

Judit Bar-Ilan *

Department of Information Science, Bar-Ilan University, Ramat-Gan, 52900, Israel and School of Library, Archive and Information Studies, The Hebrew University of Jerusalem, Jerusalem, 91904, Israel

Received 27 September 2004; accepted 3 March 2005
Available online 20 April 2005

Abstract

The Web has become an information source for professional data gathering. Because of the vast amounts of information on almost all topics, one cannot systematically go over the whole set of results, and therefore must rely on the ordering of the results by the search engine. It is well known that search engines on the Web have low overlap in terms of coverage. In this study we measure how similar are the rankings of search engines on the overlapping results.

We compare rankings of results for identical queries retrieved from several search engines. The method is based only on the set of URLs that appear in the answer sets of the engines being compared. For comparing the similarity of rankings of two search engines, the Spearman correlation coefficient is computed. When comparing more than two sets Kendall's W is used. These are well-known measures and the statistical significance of the results can be computed. The methods are demonstrated on a set of 15 queries that were submitted to four large Web search engines. The findings indicate that the large public search engines on the Web employ considerably different ranking algorithms.

Keywords: Ranking; Comparison; Search engines; Overlap

1. Introduction

The Web has become a major information source both in everyday and in professional life. Information sources available through the Web are regularly being referenced in scientific publications (Herring, 2002; Lawrence, 2001; Snyder & Peterson, 2002; Zhang, 2001). In addition, the Web is used for gathering other types of professional information (e.g. calls for papers, tables of contents and abstracts of proceedings/journal, information published by governments and different organizations).

* Tel.: +972 523667326; fax: +972 3 5353937.
E-mail address: barilaj@mail.biu.ac.il

0306-4573/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2005.03.008

The primary tools for locating information on the Web are the search engines. Currently Google is the most popular search tool and is one of the most visited sites on the Web (Nielsen/NetRatings, 2004; Sullivan, 2004a). Even though large search engines cover only a fraction of the Web (Bharat & Broder, 1998; Lawrence & Giles, 1999), for most one or two-word queries (and these are the most frequently occurring queries), the number of retrieved results is in the thousands or in the millions. Users usually browse only the first or perhaps the second page of the search results (i.e., they usually only consider the top ten or twenty results) (Silverstein, Henzinger, Marais, & Moricz, 1999; Spink, Ozmutlu, Ozmutlu, & Jansen, 2002). Thus ranking becomes crucial. In this paper we propose a new method for comparing the rankings of different search engines, based only on the documents that are listed by all the search engines that are compared.

The ranking algorithms of the search engines are not public: they are on the one hand trade secrets, and on the other hand the search engines fear that web site owners will misuse the available information in order to gain higher rankings for their pages. 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, 2004).

Since users would drown in information without reasonable ranking algorithms, it is of utmost importance to evaluate the rankings provided by different search tools. The usual method of evaluation is through human judgment, since rankings like relevance judgments are highly subjective and dependent on the context of the search carried out by the specific user. In an early study by Su, Chen, and Dong (1998), users were asked to choose and rank five most relevant items from the first 20 results retrieved for their queries. In fall 1999, Hawking, Craswell, Bailey, and Griffiths (2001) 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 (relevance was judged by humans)—a measure closely related to ranking. The results showed significant differences between the search engines. In a recent study, Vaughan (2004) 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.

For judging the rankings produced by a specific search tool, the best method is human judgment. However, for comparing rankings of different tools one can compute similarity measures without the involvement of human judges. Fagin, Kumar, and Sivakumar (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 (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. Bar-Ilan, Levene, and Mat-Hassan (2004) used three different measures to study the changes in search engine rankings over time.

We show that for the measures proposed by Fagin et al., 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 10 results of two search engines for an identical query is very small. Here we propose a different method—comparing only the comparable documents, i.e. those appearing in both lists.

2. Methodology

Before discussing in detail the methods employed by us, we briefly discuss one of the Fagin et al. (2003) metrics (all the metrics introduced by them were shown to be equivalent).

2.1. The metrics introduced by Fagin et al.

It is relatively easy to compare two rankings of the same list of items—for this well-known statistical metrics such as Kendall's tau or Spearman's footrule 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, but the extensions of Kendall's tau are shown in the paper to be equivalent. 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, 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 (which is quite plausible because of the low overlap of the search engine databases). 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. Thus a lower bound on $F^{(k+1)}(\tau_1, \tau_2)$ is:

$$F^{(k+1)}(\tau_1, \tau_2) \geqslant (k-z)(k-z+1)$$

This value is normalized by dividing the value by the maximum value of $F^{(k+1)}$, which is $k(k + 1)$, thus the normalized $F^{(k+1)}$ is a distance measure with values between 0 and 1 (0—the lists are identical). Note that the metric is heavily dominated by the non-overlapping elements, for $z = k/5$ (characteristic value we found in our experiments), the influence of the non-overlapping elements is considerable, since the minimum value of the normalized $F^{(k+1)}$ in this case will be

$$\text{normalized } F^{(k+1)} = \frac{k\left(\frac{16}{25}k + \frac{20}{25}\right)}{k(k+1)} \underset{k\to\infty}{\rightarrow} 0.64$$

Therefore in this characteristic case, the range of possible values of $F^{(k+1)}$ will be between 0.64 and 1. For the case $z = k/10$, the minimum normalized value tends to 0.81 as k goes to infinity, thus the range of possible values is even smaller.

2.2. The proposed measures

The metrics proposed by Fagin et al. (2003) can be very useful for situations where the overlap between the two compared lists is large. However, in practice, this is not the case for the search results of the large public search engines on identical queries. The major reason is that as we mentioned in the introduction the overlap of the crawled pages by the different search engines is relatively small (see Bharat & Broder, 1998; Lawrence & Giles, 1999). Thus we propose measures where we consider only the overlapping documents.

For comparing rankings of two search engines on identical queries, we compute *Spearman's rho* and test for significance. The measure is computed by considering only URLs that appear in both lists. *Spearman's rho* is applied to ranked lists of n items, where the rankings are between 1 and n. Thus the set of intersecting URLs was reranked for both search engines, where each URL received its relative rank in this subset, based on the absolute rankings of the given search engine. *Spearman's rho* ranges between −1 and 1, where 0 stands for no correlation between the two rankings, 1 for complete agreement and −1 for complete disagreement. It is the non-parametric version of Pearson's r (see for example Garson, 2004 or Lowry, 2004). For more than two lists we calculate *Kendall's W*–also called the coefficient of concordance that compares rankings of different judges (in our case different search engines) on a set of items (documents in our case). *Kendall's W* ranges between 0 (no agreement) and 1 (complete agreement) (see for example Bove, 2002). Here too we consider the set of intersecting URLs only and rerank the lists like for the case of two search engines. For both *Spearman's rho* and *Kendall's W*, one has to consider both the strength of the correlation or of the agreement (the nearer these values are to 1, the rankings are more similar) and the significance of the result, that measures whether the differences between the rankings could have been random.

Identical queries were submitted to different search engines, and the lists of all displayed results were saved. For smaller queries (i.e., less than 1000 hits) most search engines display the complete lists, while for larger queries only the first 1000 results are shown usually. We only considered search engines that provided complete rankings, thus search engines where only one or a few pages from a site are displayed, without an option to display complete lists (such options existed at the time of the data collection for Google and AltaVista) were discarded (e.g. Teoma). From the saved results pages, we filtered out the URLs of the hits and created a consolidated list of URLs that were retrieved by all the search engines for the given query. For every URL we listed the respective rank of that URL in the list of each search engine in which the URL appeared. For every pair of search engines we looked at those URLs only that appeared in both lists and reranked the URLs appearing in the intersection.

The Web search engine scene is rather dynamic, since the data collection AlltheWeb and AltaVista have undergone major changes, currently both are powered by Yahoo (Sullivan, 2004b). Thus, had we run these queries in October 2004, we probably would have received different results, however, the methods presented here are still valid, and based on smaller scale experiments we conclude that there are still huge differences between the ranking algorithms of Yahoo and Google. Search engines change their ranking algorithms all the time (see for example the reports on the Florida Google dance—Sullivan, 2003) even without changes in their business models, thus in any case the specific results presented here were only valid at the time of the data collection. The specific results presented here demonstrate the applicability of the methods, and serve as an alert as to the huge differences in ranking by the different systems.

2.3. Data collection

For our experiment we chose 15 queries in the area of information retrieval. The queries were not chosen randomly, since we wanted to experience with different sizes of results sets, based on the number of results reported by Google—the first five queries resulted in less than 200 hits, for the next five the number of hits was between 200 and 1000, and the last five queries had more than 1000 hits—for this set only the first 1000 results were retrieved (1100 for AlltheWeb). We queried four search engines: Google, AlltheWeb, AltaVista and HotBot on December, 7, 2003. For Google we used the "the repeat the search with the omitted results included" option, and for AltaVista we marked the "site collapse off" option in order to have complete rankings of the retrieved documents. Table 1 displays the queries and the number of URLs retrieved for each query by each search engine separately and by all the search engines together. For the large queries we also show the number of reported results in parentheses. In brackets is the relative coverage of each search engine in percentages out of the pool of URLs identified by the four search engines for the query.

For the small and medium queries the relative coverage of Google is considerably higher that those of the other search engines, while in December 2003, HotBot had the lowest coverage. As pointed out before, these results only present a momentary snapshot, and no conclusions should be drawn from the specific findings—they are of anecdotal value only.

Table 1
The queries and the number of retrieved results and the total number of different URLs identified for the query, and the coverage of each search engine out of the pool of URLs identified for the query (in brackets)

Query no.	Query	Google	AlltheWeb	AltaVista	HotBot	Total
q01	"relative ranking" TREC	58	30	24	6	88
		[66%]	[34%]	[27%]	[7%]	
q02	"SIGIR 2004"	177	40	27	30	219
		[81%]	[18%]	[12%]	[14%]	
q03	"everyday life	170	114	60	67	262
	information seeking"	[65%]	[44%]	[23%]	[26%]	
q04	"natural language processing for IR"	137	78	12	63	229
		[60%]	[34%]	[5%]	[28%]	
q05	"multilingual retrieval"	22	6	4	4	30
	Hebrew	[73%]	[20%]	[13%]	[13%]	
q06	"relative ranking" IR	568	292	114	56	792
		[72%]	[37%]	[14%]	[7%]	
q07	"social network analysis"	748	408	169	206	1120
	"information retrieval"	[67%]	[36%]	[15%]	[18%]	
q08	"link mining"	269	209	115	84	464
		[58%]	[45%]	[25%]	[18%]	
q09	"citation analysis"	377	157	65	57	652
	"link structure"	[58%]	[24%]	[10%]	[9%]	
q10	bibliometrics	315	106	60	46	382
	"link analysis"	[82%]	[28%]	[16%]	[12%]	
q11	relevance ranking	1000 (1640)	696 (1233)	426 (708)	345 (540)	1791
	"link analysis"	[56%]	[39%]	[24%]	[19%]	
q12	"Cross-language retrieval"	1000 (2480)	1093 (5675)	1000 (1100)	421 (840)	2477
		[40%]	[44%]	[40%]	[17%]	
q13	"Question answering" IR	1000 (3960)	1088 (8346)	1000 (1177)	671 (1492)	2709
		[37%]	[40%]	[37%]	[25%]	
q14	"information retrieval"	1000 (1,090,000)	1100 (1,091,278)	994 (264,826)	449 (247,992)	2914
		[34%]	[38%]	[34%]	[15%]	
q15	"data mining"	998 (1,730,000)	1100 (1,274,549)	963 (367,411)	450 (384,846)	2856
		[35%]	[39%]	[34%]	[16%]	

3. Results and discussion

We computed *Spearman's rho* between every pair of search engines and for every query. The results appear in Table 2: in the table one can find the significance of the results and the size of the overlap for every pair. We also calculated the average correlation between pairs of search engines, based on the overlapping URLs.

Interesting to note that even though the highest correlation was between Google and AltaVista, the correlation was especially high (and in most of the cases significant) for most of the queries, there were a few exceptions, especially for query 11, 'relevance ranking AND "link analysis"'—here the correlation was extremely low, only 0.004, even though both engines indexed 128 common documents. Another point to notice is that in spite of the low correlation between the rankings of Google and AlltheWeb, they both correlated relatively strongly with AltaVista.

For the sake of comparison we also computed F^{51} measure introduced by Fagin et al. (2003) for queries 6–15 (for the first queries there were less than 50 hits for some of the search engines). The results were normalized so the numbers are between 0 and 1 (all values were divided by 2550—the highest possible value), just like in (Fagin et al., 2003). The results appear in Table 3 together with the size of the overlap (the z value in the formula for computing $F^{(k+1)}$) for the two top 50-lists. Note that for query 10, HotBot retrieved only 46 results.

Table 2
Correlations between pairs of search engines on the queries

	Google–Alltheweb			Google–AltaVista			Google–Hotbot			AlltheWeb–AltaVista			AlltheWeb–HotBot			AltaVista–Hotbot		
	Sp	sig	ol	Sp	sig	ol	Sp	sig	ol	Sp	sig	ol	Sp	sig	ol	Sp	sig	ol
q01	0.6	0.85	13	0.8	*0.00*	19	1.0	*0.00*	2	0.4	0.40	6	0.4	0.60	4	1.0	*0.00*	1
q02	−0.4	**0.02**	26	0.7	*0.00*	19	0.7	*0.00*	17	0.8	**0.02**	9	−0.4	0.31	10	0.3	0.49	9
q03	−0.4	**0.02**	26	0.7	*0.00*	19	0.7	*0.00*	17	0.8	**0.02**	9	−0.4	0.31	10	0.3	0.49	9
q04	−0.4	0.78	44	0.1	0.78	8	0.4	**0.03**	44	0.7	0.19	5	0.5	0.06	16	0.1	0.87	5
q05	1.0	*0.00*	2	0.5	0.67	3	1.0	*0.00*	2	1.0	*0.00*	2	−1.0	1.00	2			0
q06	0.0	0.69	125	0.7	*0.00*	67	0.4	0.05	22	0.2	0.13	42	0.1	0.72	28	0.4	0.20	12
q07	0.0	0.94	69	0.8	*0.00*	80	0.4	**0.03**	31	0.1	0.34	59	0.1	0.45	32	−0.1	0.57	22
q08	0.0	0.94	69	0.8	*0.00*	80	0.4	**0.03**	31	0.1	0.34	59	0.1	0.45	32	−0.1	0.57	22
q09	0.0	0.79	100	0.6	*0.00*	49	0.5	**0.01**	28	0.2	0.27	34	0.0	0.98	28	0.5	**0.04**	18
q10	0.1	0.23	66	0.6	*0.00*	46	0.3	0.24	22	0.4	0.08	23	0.3	0.35	16	−0.4	0.43	7
q11	0.0	0.89	133	0.0	0.96	128	−0.1	0.42	78	0.3	*0.00*	162	0.2	**0.02**	146	0.4	*0.00*	108
q12	0.3	*0.00*	258	0.3	*0.00*	263	0.4	*0.00*	99	0.5	*0.00*	412	0.3	*0.00*	163	0.4	*0.00*	160
q13	0.1	0.44	220	0.5	*0.00*	270	0.4	*0.00*	100	0.3	*0.00*	341	0.1	0.05	191	0.3	*0.00*	159
q14	0.7	*0.00*	177	0.7	*0.00*	225	0.6	*0.00*	85	0.7	*0.00*	177	0.6	*0.00*	96	0.5	*0.00*	90
q15	0.6	*0.00*	178	0.7	*0.00*	193	0.4	*0.00*	77	0.6	*0.00*	247	0.4	*0.00*	92	0.5	*0.00*	97
Average	**0.14**	0.44	100.4	**0.56**	0.16	97.9	**0.50**	0.05	43.7	**0.47**	0.12	105.8	**0.10**	0.35	57.7	**0.29**	0.26	47.9

'Sp' stands for Spearman's rho; 'sig' for significance (significant at the 0.05 level is bolded, and significant at the 0.01 level is in italics; ol stands for the size of the overlap).

Table 3
Normalized F^{51} for queries 6–15

	Go–All	Overlap	Go–Alt	Overlap	Go–Hot	Overlap	All–Alt	Overlap	All–Hot	Overlap	Alt–Hot	Overlap
q06	0.933	4	0.801	12	0.818	7	0.855	8	0.884	7	0.816	10
q07	0.958	3	0.920	8	0.934	5	0.899	7	0.910	6	0.889	8
q08	0.963	2	0.362	34	0.964	4	0.915	8	0.874	9	0.968	3
q09	0.947	3	0.785	9	0.916	4	0.769	11	0.788	12	0.707	17
q10	0.968	2	0.838	7	0.928	3	0.755	12	0.794	11	0.864	7
q11	0.967	1	1.000	0	1.000	0	0.846	8	0.889	6	0.811	8
q12	0.962	3	0.973	6	0.962	3	0.804	10	0.912	5	0.951	3
q13	0.955	2	0.884	9	0.917	5	0.857	5	0.947	4	0.924	6
q14	0.744	12	0.627	23	0.791	10	0.751	10	0.813	9	0.872	9
q15	0.776	13	0.673	21	0.910	5	0.540	24	0.804	9	0.871	7
Average	**0.917**	4.5	**0.786**	**12.9**	**0.914**	4.6	**0.799**	10.3	**0.861**	7.8	**0.867**	7.8

F is a distance measure, therefore, the smaller the value the more similar are the rankings of the compared search engines. Let us compare the rankings based on the average of *Spearman's rho* with the ranking based on the *F* metric. The comparison appears in Table 4.

Table 4 shows differences in rankings of similarity/correlation. The differences between the *F* values are small, compared to the far more emphasized differences in the average values for *Spearman's rho*. *Spearman's rho* is a well-known measure, and categorizations of the strength of the correlation exist. Some say that if the absolute value of the correlation is below 0.3 then the correlation is weak, between 0.3 and 0.49 it is medium and above 0.5 it is large (Cohen, 1988). Others define the correlation as strong, only if the absolute value is above 0.7, and medium when the absolute value is between 0.4 and 0.7 (Rowntree, 1981). In either case we see that the average correlation between Google and AltaVista and Google and Hotbot is medium-high, while the correlations between Google and AlltheWeb and AlltheWeb and Hotbot are very low.

The most striking difference between the two rankings is the placement of the Google-HotBot pair. We believe that the *F* measure is very strongly influenced by the size of the overlap, as we discussed in the methodology section. Actually the rankings for the *F* measure are exactly according to the size of the average overlap. Thus it seems that the *F* measure is useful in situations where the overlap between the results of every two pairs of compared search tools is considerable. *Spearman's rho*, on the other hand, totally ignores the non-overlapping elements, and concentrates only on the differences in the rankings of the search engines being compared.

We also compared groups of three search engines on the set of URLs that were retrieved by all three engines. Finally we looked at the relative rankings of the URLs retrieved by all four search engines. For the case of three or more tools, we computed *Kendall's W* that measures the agreement between raters and ranges between 0 and 1: 0 for no agreement and 1 for complete agreement. SPSS provides a significance measure associated with this value. The *df* in the table stands for degrees of freedom and it is the size of the intersection minus 1. The results appear in Table 5.

Table 4
Comparing the correlation data with the F^{51} metric

Rank	Based on correlation	Average Spearman	Based on F^{51}	Average F^{51}
1	Google–AltaVista	0.563	Google–AltaVista	0.786
2	Google–HotBot	0.504	AlltheWeb–AltaVista	0.799
3	AlltheWeb–AltaVista	0.475	AlltheWeb–HotBot	0.861
4	AltaVista–HotBot	0.285	AltaVista–HotBot	0.867
5	Google–AlltheWeb	0.138	Google–HotBot	0.914
6	AlltheWeb–HotBot	0.098	AlltheWeb–HotBot	0.917

Table 5
Agreement on the rankings between sets of three or more search engines

	Go–All–Alt			Go–All–Hot			Go–Alt–Hot			All–Alt–Hot			Go–All–Alt–Hot		
	W	df	sig	*W*	df	sig	*W*	df	sig	*W*	df	sig	*W*	df	sig
q01	0.72	5	0.055	0.39	1	0.083		0			0			0	
q02	0.39	7	0.313	0.24	8	0.682	0.52	5	0.170	0.47	3	0.241	0.18	3	0.552
q03	0.38	31	0.256	0.35	25	0.397	0.22	12	0.819	0.51	12	0.095	0.19	12	0.675
q04	0.58	4	0.139	0.76	13	*0.005*	0.64	3	0.122	0.78	2	0.097	0.81	2	**0.039**
q05		0			0			−1			−1			−1	
q06	0.58	33	*0.006*	0.57	12	**0.056**	0.62	9	0.053	0.59	5	0.113	0.59	5	**0.037**
q07	0.47	58	**0.022**	0.40	58	0.151	0.60	34	*0.003*	0.44	26	0.132	0.38	23	0.051
q08	0.55	47	*0.003*	0.51	16	0.080	0.62	14	*0.025*	0.49	12	*0.110*	0.63	11	*0.004*
q09	0.45	26	0.116	0.49	18	0.096	0.62	13	**0.029**	0.61	14	**0.029**	0.52	11	**0.017**
q10	0.64	17	**0.012**	0.38	10	0.323	0.33	4	0.406	0.24	5	0.613	0.23	4	0.463
q11	0.37	65	0.238	0.33	45	0.501	0.36	43	0.353	0.57	65	*0.000*	0.27	29	0.336
q12	0.61	155	*0.000*	0.52	56	*0.004*	0.52	57	*0.004*	0.59	92	*0.000*	0.50	38	*0.000*
q13	0.48	125	*0.001*	0.46	51	*0.034*	0.60	58	*0.000*	0.45	89	*0.015*	0.41	37	*0.010*
q14	0.78	104	*0.000*	0.79	51	*0.000*	0.70	53	*0.000*	0.73	52	*0.000*	0.74	42	*0.000*
q15	0.78	114	*0.000*	0.64	43	*0.000*	0.66	50	*0.000*	0.66	57	*0.000*	0.64	38	*0.000*
Average	**0.56**	**52.7**	0.08	**0.49**	**27.1**	0.17	**0.54**	**23.6**	0.15	**0.55**	**28.9**	0.11	**0.47**	**16.9**	0.17

W stands for Kendall's *W*; sig for significance (significant at the 0.05 level is bolded, and significant at the 0.01 level is in italics; df stands for degrees of freedom (size of intersection minus 1)).

We see that the average agreement for all the sets is around 0.5, and the results for the larger queries are mostly significant, i.e. the outcome is with high probability not accidental. For query 5, "multilingual retrieval" AND Hebrew, there was not a single document that was retrieved by all four search engines.

4. Conclusions

The findings of this study clearly show that search engines employ very different ranking algorithms. The obvious differences in rankings methods should be noted by experienced users, especially by scientists looking for a wide range of quality information on the Web. By submitting the same query to several search tools or to a meta search engine that sends the query to the major search engines, even by looking only at the top-10 or top-20 results retrieved by each of the search tools, one can increase the range of the results considerably.

Our method of taking into account only documents that are retrieved by all the search engines that are being compared, zooms in on the differences in the "ranking recipes", without penalizing the search engines for having smaller indices or for crawling in places that the other search engines have not reached.

The method presented here compares rankings of different search engines; however it says nothing about the superiority of one ranking over the other. In order to decide which ranking is better in the "eyes of the users", large-scale user studies have to be carried out along the lines of Vaughan's research (2004).

As for the future, we feel that larger scale studies are appropriate. Such studies should be carried out periodically, since both the Web and the ranking algorithms undergo constant changes.

References

Bar-Ilan, J., Levene, M., & Mat-Hassan, M. (2004). Dynamics of search engine rankings—a case study. In *Proceedings of the 3rd international workshop on web dynamics, New York, May 2004*. Available: <http://www.dcs.bbk.ac.uk/webDyn3/webdyn3_proceedings.pdf>.

Bharat, K., & Broder, A. (1998). 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, pp. 379–388). Available: <http://decweb.ethz.ch/WWW7/1937/com1937.htm>.

Bove, R. E. (2002). Correlation. Available: <http://courses.wcupa.edu/rbove/eco252/252corr.doc>.

Cohen, J. (1988). *Statistical power analysis for behavioral sciences*. Hilldale, NJ: Erlbaum.

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

Garson, D. (2004) Correlation. In *Qualitative methods in public administration*. Available: <http://www2.chass.ncsu.edu/garson/pa765/correl.htm>.

Google (2004). Information for Webmasters. Available: <http://www.google.com/webmasters/4.html>.

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

Herring, S. D. (2002). Use of electronic resources in scholarly electronic journals: A citation analysis. *College & Research Libraries, 63*(4), 334–340.

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature, 411*, 521.

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

Lowry, R. (2004). Rank-order correlation. In *Concepts and applications of inferential statistics*. Available: <http://faculty.vassar.edu/lowry/ch3b.html>.

Nielsen/NetRatings. (2004). NetView usage metrics. Available: <http://www.netratings.com/news.jsp?section=dat_to>.

Rowntree, D. (1981). *Statistics without tears: A primer for non-mathematicians*. Penguin.

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

Snyder, S. J., & Peterson, A. (2002). The referencing of internet web sites in medical and scientific publications. *Brain and Cognition, 50*, 335–337.

Spink, A., Ozmutlu, S., Ozmutlu, H. C., & Jansen, B. J. (2002). US versus European Web searching trends. *SIGIR Forum Fall*, Available: <http://www.acm.org/sigir/forum/F2002/spink.pdf>.

Su, L. T., Chen, H. L. & 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* (Vol. 35, pp. 348–361).

Sullivan, D. (2003). Florida Google dance resources. Available: <http://www.searchenginewatch.com/searchday/article.php/3285661>.

Sullivan, D. (2004a). Nielsen NetRatings search engine rankings. In *Searchenginewatch reports*. Available: <http://searchenginewatch.com/reports/article.php/2156451>.

Sullivan, D., (2004b). Who powers whom? Search providers chart. In *Searchenginewatch reports*. Retrieved October 15, 2004, from Available: <http://searchenginewatch.com/reports/article.php/2156451>.

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

Zhang, Y. (2001). Scholarly use of Internet-based electronic resources. *Journal of the American Society for Information Science and Technology, 52*(8), 628–650.

Judit Bar-Ilan is currently a senior lecturer at the Department of Information Science at Bar-Ilan University in Israel. She is also a faculty member of the School of Library, Archive and Information Studies at the Hebrew University of Jerusalem. Her research interests include Internet research, information retrieval, informetrics and information behavior.