



SAVVY SEARCHING

Relevance in the eye of the search software

Péter Jacsó

University of Hawaii, Manoa, Hawaii, USA

Abstract

Purpose – The purpose of this article is to look into relevance ranking and its importance in trying to bring some order to the deluge of results in response to a query.

Design/methodology/approach – A large-scale analysis of detailed web logs of various search engines was performed. Sample tests were made on five to eight versions of MEDLINE, ERIC, and PsycINFO on hosts which have comparable versions of the databases and offer relevance ranking.

Findings – It was found that, for fairness, it must be ensured that the implementations are identical, they have the same retrospective coverage, the same MEDLINE/PubMed subsets, and (quasi) identical update.

Research limitations/implications – The tests were made early September 2005. As databases are updated at different times, perfect synchronicity is not easy to achieve. When new records are added to the database, they may change the ranking of the test result set. Similarly, a small change in the fine-tuning of the algorithm may yield different rank order positions of the same record the next time.

Originality/value – Brings together important research findings and suggests a topic for the next column.

Keywords Information retrieval, Search engines, Databases, Computer software

Paper type Technical

Relevance of documents retrieved has been the cardinal criterion in evaluating information retrieval systems. For the two most important measures of the effectiveness of search process, recall and precision, relevance is used in calculating the recall and precision scores. All the web-wide search engines use relevance ranking in ordering the results presented to the users – if for nothing else, out of necessity. Relevance has been considered the most objective measure to gauge how effective the retrieval system is in finding documents related to the topic represented by the query. This type of relevance is also referred to as algorithmic or statistical relevance, which does not consider the searchers' non-topical preferences for currency, document type, etc. Traditional commercial information retrieval systems have sorted and presented results by date, author, journal name, author affiliation and some other less common database-specific data elements. Most of them now also offer relevance-ranked result lists, which are supposed to order and present the hits by decreasing probability of relevance to the query submitted as decreasingly best matches. Sample tests using the same databases on different hosts show significant differences in the relevance-ranked result lists for functionally identical queries. This indicates a lack of consensus among the search systems in determining the topical relevance of the same documents or document surrogates in the same database context.



The profile of the average web searcher

To understand why relevance ranking is critical for trying to bring some order to the deluge of results in response to a query, one needs to know the features of a typical user search. Greisdorf and Spink (2000) reported on the importance and highlights of relevance research, with special attention to their implications for information professionals, i.e. the savviest searchers. Their paper includes a very good bibliography on the topic for further reading.

Large-scale analyses of detailed web logs of various search engines have found highly characteristic patterns in searchers' behaviour. The findings are invaluable for their reproducibility as well as for the size of the population. The study of Spink *et al.* (2001) had details for 1.2 million queries; Silverstein *et al.* (1999), for close to one billion queries; and Xu (1999), for a staggering 30 billion queries. The log data provided information about the search terms and operators used, the number of results pages viewed, the query reformulations steps, and other transactions for all the queries in the time frames analysed. The Spink study was a follow-up to an earlier, relatively small study of 51,000 queries by Jansen *et al.* (2000). These population-wide measures are far more informative than the small samples typically used until the late 1990s for observing and analysing the information-seeking behaviour of users.

The most important measures from the point of view of relevance ranking include sobering facts about the typical searcher. The average number of words per queries was 2.35 both in the Silverstein *et al.* (1999) and the Spink *et al.* (2001) analyses. According to the latter, 80 per cent of the users did not use any search operators, such as the Boolean "and", "or", "not", or their functionally equivalent symbols, or pull-down menu options. It also reported that 95 per cent of users did not use the quotation marks for phrase searching. It was also discovered that more than 60 per cent of search sessions consisted of a single request and the display of a single page of the result list – typically with ten summaries of results/screen. This was the median value with a mean of 1.38 screens (typically with summaries of ten results, the default value for items/screen). In the Silverstein *et al.* (1999) analysis this number was higher with a mean of 2.21 screens per query.

In any interpretation these statistics clearly indicate that few users look beyond the first screen, or bother to formulate the query with Boolean, proximity and positional operators, let alone reformulate the query. Although there are no statistics about the use of non-topical limits (filters) in the very large query sets, it is safe to say that they are negligible, as often they appear on the advanced search template, which are very rarely used according to anecdotal, but very consistent, evidence by insiders in the information industry. Limiting the query to the title, descriptor field or major descriptors certainly represents smart search strategy, but the distinction between major and minor descriptors is not entirely clear even for many librarians. In addition the feature is used by only a few databases, and even then the filtering option may not be available in all of the implementations of the database. CSA, for example, does not offer this option. Even from a much smaller result set, such as the one about medical informatics yielding about 550 records, it is an arduous process to wade through the list to find the most likely relevant articles based on the few data elements which appear in the short result lists (title, author, publication year). Relevance ranking will come to the rescue. Or will it?

The relevance-ranking guessing game

The result lists sorted by author, publication year and/or journal name are easy to understand, as they are quite transparent. If there is some idiosyncrasy in the list, it stands out. Sorting (ranking) by relevance is enigmatic at best. That is the reason why it received only a lukewarm mention in the review of output options in Jacsó (2005), emphasising that the rank orders may be questionable, and the process is not verifiable because the details of the ranking algorithms are not revealed beyond generalities.

Only a few systems provide some hints about the absolute relevance score, such as KnowledgeFinder and DIALOG (if you use the TARGET function in command mode). The calculation of the relevance score is still not transparent. From the sample in Figures 1 and 2 it is clear that the number of occurrences of the exact phrasal term, “medical informatics”, is not the only criterion for calculating the relevance score of a document (surrogate). In the first record the search term appears once in the title, five

The screenshot shows the DialogWeb search interface. At the top, there is a navigation bar with icons for search, favorites, settings, order, cost, logoff, and help. Below this, the search criteria are displayed: 'Dynamic Search: MEDLINE® (1951-present)' and 'Picklist for: medical informatics'. The results summary indicates '5117 records found. 50 displayed. [Display all records unsorted]'. The interface includes controls for 'Output' (show rates...), 'Modify' (sorted by: Relevance), and 'format' (Full Record). A 'select' dropdown is set to 'all'. The results list shows four records, each with a checkbox and a title: 1. Medical informatics: once more towards systematization, Sep 1996, MEDLINE(R) (File 155); 2. The internal challenges of medical informatics, Mar 1997, MEDLINE(R) (File 155); 3. Health and medical informatics education, Jun 1994, MEDLINE(R) (File 155); 4. The medical informatics curriculum at the University of Heidelberg/School of Technology.

Figure 1.
Excerpt from relevance ranked output of Dialog Web

DIALOG-TARGET RESULTS (arranged by percent RELEVANCE)

```

----- Item: 1 -----
*Medical *informatics.: once more towards systematization.
- Statistical Relevance: 99
- Term Frequency: MEDICAL INFORMATICS - 24
----- Item: 2 -----
The internal challenges of *medical *informatics..
- Statistical Relevance: 94
- Term Frequency: MEDICAL INFORMATICS - 28
----- Item: 3 -----
Health and *medical *informatics..education.
- Statistical Relevance: 84
- Term Frequency: MEDICAL INFORMATICS - 28

```

Press ENTER to continue browsing or enter item number(s) to see full record
M = Modify search I = New TARGET C = Customize display Q = QUIT

Figure 2.
Some details about the term frequency and relevance score

times in the abstract, and twice as a MeSH subject heading - once as a major one. (The term frequency is there, but it is obscure and does not reflect what we see in the record.) The pattern is the same for the record ranked as Number 2, but its relevance score is only 94. The length of the two records is similar. The record ranked third has the same pattern except for the single occurrence of “medical informatics” as a MeSH term, but that one is in the capacity of major heading. The term frequency as reported by Dialog is the same as for the second ranked record, but the relevance score is merely 84. Although the calculation of the relevance score and the term “frequency” is obscure, it is still more than most other systems would reveal.

Comparative rank analysis

Except for one option discussed below, there is not much even knowledgeable searchers could explore about the credibility and mechanism of objective relevance ranking. Their own judgments about the relevance of the retrieved items would be subjective, and implicitly include non-topical aspects. Furthermore, they would not know how many additional, more relevant items there are in the set retrieved unless they uncover and evaluate all of them. This is not feasible, as even in the smaller databases and with seemingly esoteric topics there may be thousands of records in the set.

The one option is to run the functionally identical search in several implementations of the same database on different host systems. For fairness, it must be ensured that the implementations are identical, they have the same retrospective coverage, the same MEDLINE/PubMed subsets, and (quasi) identical update. Take MEDLINE as the most widely implemented database. It has versions which have only a few years of retrospective coverage, or do not have the most currently created update records, do not include the in-process and/or old (pre-1996) Medline records, or the additional PubMed records. The differences among the retrieval software must be compensated for by the searcher to ensure a level playing field. For example, Ebsco does automatic pluralisation and singularisation for many words and also British/American transformation for most of the differently spelled words (e.g. “organisation” and “organization”, but not for “encyclopedia” and “encyclopaedia”). It is still worth the effort to normalise the queries, as significant rank order differences of identical or almost identical result sets can be illuminating for searchers who do care about the issue, who often need to discover the most relevant 10-30 documents in different subject fields, and not merely find a good enough article for the next assignment.

The tests were made early September, 2005. As databases are updated at different times, perfect synchronicity is not easy to achieve. When new records are added to the database (even if not related to the topic of the test search), they may change the ranking of the test result set. Similarly, a small change in the fine-tuning of the algorithm may yield different rank order positions of the same record the next time.

Sample tests were made on 5-8 versions of MEDLINE, ERIC, and PsycINFO on hosts which have comparable versions of the databases and offer relevance ranking. For example CSA is excluded from MEDLINE tests, as it has a relatively small subset of this large database. So are Ovid, PubMed and DIMDI, which do not offer relevance ranking. (Ovid has relevance ranking for the Books@Ovid database).

Queries were used which produce relatively small result sets from the databases (about 30 to 150 records) to keep the comparative evaluation feasible for this one-man operation, but still produce telling examples for the level of congruence in the ranking

of the result sets. The limits on the number of relevance ranked items made this restriction on size also necessary. Dialog's limit is 50 records; KnowledgeFinder's is 200. In this column space allows only a very few black-and-white illustrations, but additional colour figures are available at: www2.hawaii.edu/~jacso/savvy/relevance

Telling example

One of the simplest tests was about the topic of scientometrics (using the query "scientometric OR scientometrics", or its equivalent), which yielded the same 31 hits from four MEDLINE implementations. However, the rank order position of the same document showed a low consensus among these four hosts.

As Figure 3 illustrates, only a few of the 31 items were ranked close enough by all four or at least by three systems shown here, such as in the case of Records 2, 7, 9, 16 and 17. These four implementations had a complete match. Adding the relevance rank position of the other MEDLINE implementations would show a far larger distance among the rank positions.

A more informative visual representation of the scatter is presented in Figure 4 on the following page. The web page (www2.hawaii.edu/~jacso/savvy/relevance) has a colour version of the figure, which helps in identifying the pattern of pairs of hosts with the closest consensus in ranking the items. Other tests for result sets of almost identical

Rank order position				Item Id	Title
Wok	Sci	OCL	Ebs		
13	8	30	13	1	[A epistemometric view of some biological dis
22	23	24	23	2	[A scientometric analysis of the literature on r
7	15	18	18	3	[A scientometric radiography of Revista Medi
27	18	13	27	4	[Biological and scientometric characteristics d
10	5	8	4	5	[Impact factors and bibliometrics of science.
31	24	6	29	6	[Importance of scientometrics and bibliometry
21	25	22	21	7	[Interdisciplinary research in gerontology: cita
25	21	28	26	8	[Is the medical research carried out in Croatia
23	19	23	22	9	[Overview of research in Chile by several epis
18	27	11	15	10	[Scientific literature: bibliometric and bibliogra
24	29	27	25	11	[Scientific-metric analysis of structure and ag
26	31	29	28	12	[Scientometric analysis of modern trends in th
1	10	4	9	13	[Scientometric and publication malpractices.
30	20	5	31	14	[Scientometric approach to studying trends in
28	17	16	30	15	[Scientometric characteristics of the publishe
4	2	3	5	16	[Scientometrics and bibliometrics of biomedic
3	1	2	6	17	[Scientometrics and bibliometrics of chronoth
5	9	1	10	18	[Scientometrics and publishing in Hungarian
15	7	21	11	19	[Scientometrics of medical journals in the fore
16	13	25	19	20	[The developmental trend in research on radi
6	12	15	7	21	[The evaluation of research performance in p
17	14	26	17	22	[The impact factor--a reliable sciento-metric p
11	3	9	3	23	[The journal impact factor as a parameter for
19	30	12	14	24	A soft systems approach to designing an info
14	6	31	12	25	Assessing oncological productivity. is one me
12	4	10	2	26	Changes in the basic experimental paramete
20	26	14	20	27	Citation analysis of 541 articles published in c
9	22	20	24	28	Quantitative analysis of current trends in the
8	16	19	16	29	Scientometric analysis of anthropology in the
29	28	17	32	30	Scientometric study of Health Physics.
2	11	7	8	31	The use of scientometric parameters for the

Figure 3.
Rank order positions of the same MEDLINE records from the result lists of four hosts

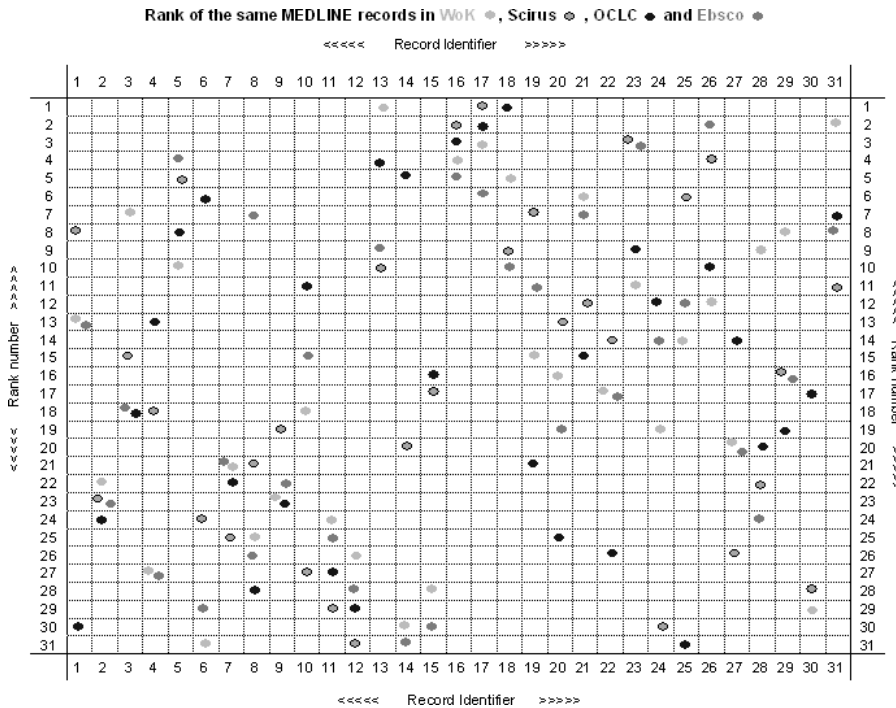


Figure 4.
Scatter of rank positions
for the same MEDLINE
records in four
implementations

or larger size showed a similar pattern of scatter. In cases where an implementation was not current enough, and/or did not include MEDLINE records with in-process status, the ranking of the query results showed a much wider scatter.

The picture, however, may not be as discouraging as it looks. The distance between rank positions within a single system is not linear as the sequence numbers in the result list may suggest. The common solution in ranked lists is to assign the same rank position to two or more items with identical or very close rank scores, declaring a tie. However, this does not work with search results. The unique sequence numbers are needed to mark records individually for more detailed display, saving or printing. It certainly happens that there is no difference, or there is minimal difference among the relevance scores of the items ranked from, say, Record 3 to 8 in relevance ranked search results.

Looking at the details of DIALOG's TARGET lists clearly proves this by showing the relevance score and the occurrence of the search terms, which inform the searchers – although in a rather clunky format. The best solution would be to show the relevance score next to each sequence number in the short result list to inform the searchers of the real rank order difference among the items retrieved.

The proof of the pudding is in the eating, of course. Savvy searchers pick up subtle signs of non-topical relevance clues, such as the brackets around the title of many items on the above result list, which indicate that the articles are not in English. The (ir)relevance of the items as judged by the users from their perspective can be very

different from the one suggested by the objective, purely statistics-based relevance ranking algorithm. That is the topic for the next column.

References

- Greisdorf, H. and Spink, A. (2000), "Recent relevance research: implications for information professionals", *Online Information Review*, Vol. 24 No. 5, pp. 389-95.
- Jacsó, P. (2005), "Options for presenting search results: Part I: Common options", *Online Information Review*, Vol. 29 No. 3, pp. 311-19.
- Jansen, B.J., Spink, A. and Saracevic, T. (2000), "Real life, real users and real needs: a study and analysis of users' queries on the web", *Information Processing and Management*, Vol. 36 No. 2, pp. 207-27.
- Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. (1999), "Analysis of a very large web search engine query log", *ACM SIGIR Forum*, Vol. 33 No. 3, pp. 6-12.
- Spink, A., Wolfram, D., Jansen, B.J. and Saracevic, T. (2001), "Searching the web: the public and their queries", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 3, pp. 226-34.
- Xu, J. (1999), "Internet search engines: real world IR issues and challenges", paper presented at the Conference on Information and Knowledge Management, CIKM 99, October 31-November 4, Kansas City, MO.