Jointly published by Elsevier Science Ltd, Oxford
and Akadémiai Kiadó, Budapest

Scientometrics,
Vol. 42, No. 2 (1998) 207–228

ON THE OVERLAP, THE PRECISION AND ESTIMATED RECALL OF SEARCH ENGINES. A CASE STUDY OF THE QUERY "ERDOS"

JUDIT BAR-ILAN

School of Library, Archive and Information Studies, The Hebrew University of Jerusalem,
Jerusalem, 91904 (Israel)

(Received January 22, 1998)

In this paper we investigate the retrieval capabilities of six Internet search engines on a simple query. As a case study the query "Erdos" was chosen. Paul Erdos was a world famous Hungarian mathematician, who passed away in September 1996. Existing work on search engine evaluation considers only the first ten or twenty results returned by the search engine, therefore approximation of the recalls of the engines has not been considered so far. In this work we retrieved all 6681 documents that the search engines pointed at and thoroughly examined them. Thus we could calculate the precision of the whole retrieval process, study the overlap between the results of the engines and give an estimate on the recall of the searches. The precision of the engines is high, recall is very low and the overlap is minimal.

1. Introduction

Recently we are experiencing the "Web document explosion". There are already more (maybe much more) than a hundred million Web pages. In December 1997, the search engine AltaVista[*] announced that it indexed 100 million Web pages. *Feldman* (1997a) reports that Louis Monier, Chief Technology Officer of AltaVista, guesses that search engines index only half of the total number of Web pages. If his guess is correct, there are already about 200 million pages on the Net. This number seems to continue to grow exponentially. With such a vast amount of information on the Net there is an acute need for tools to help us find a needle (or needles) in a haystack. The search engines are one of the main finding aids on the Web. But do the existing search engines fulfill this job? Is their *recall* and *precision* satisfactory? Is there one "best" search engine?

0138–9130/98/US $ 15.00

Lancaster and *Fayen* (1973) listed six criteria for evaluating on-line retrieval systems: recall, precision, response time, user effort, form of output and coverage. In this paper we address the first two criteria applied to the Web on a given search query. In addition we analyze the overlap between the search engines and study other characteristics of the retrieved documents.

A survey of previous works on the subject shows that the papers can be partitioned into two groups. The first group evaluates, describes and compares engines. *Dong & Su* (1997) present a thorough review article on search engine evaluation with extensive reference to previous work. They found that "three types of methodologies have been used in evaluating search engines: actual tests with data collection and analysis; evaluative comments with examples of simple searches; and review of functions of different searches without examples". The two main measures of retrieval effectiveness, precision and recall, have not received much attention. They state that recall is a difficult measure to apply on the Web, and thus has been abandoned. In previous work, precision has been calculated but only for the first ten or twenty retrieved results, see for example (*Tomaiulo & Packer*, 1996), (*Chu & Rosenthal*, 1996), (*Ding & Marchionini*, 1996), (*Leighton & Srivastava*, 1997), (*Venditto*, 1996), (*Feldman*, 1997b) and (*Haskin*, 1997). *Zorn* et al. (1996) report on the number of results of three complex queries submitted to four search engines along with some remarks on the relevance of the results. They say that "the conclusion was reached that no single Web search system is really 'the best'. None of the four systems can claim to include *all* of the Internet in their databases". There is no supporting evidence for this claim. *DeZeler-Tiedman* (1997) carried out a different kind of study: known-item searching of the titles of thirty nine Web pages, looking for their URLs using four search engines. The results were quite satisfactory.

The second group of papers use bibliometric methods to study Web documents, but they all follow different directions from the one taken in this paper. A very large case study was conducted by *Woodruff* et al. (1996), they examined 2.6 million Web documents for characteristics like document size and automatically analyzed 92,000 documents for html syntax errors. Other studies ((*Larson*, 1996), (*Bar-Ilan*, 1997), (*Almind & Ingwersen*, 1997), and (*Rousseau*, 1997)) showed the applicability of informetric methods and functions to the Internet.

As one can see from the literature survey, as of today, there are no search engine evaluations that analyze the whole list of URLs returned by the search engines. In this paper we concentrate on a single simple query, thoroughly analyze and compare the search results of seven search engines. We examine the retrieved documents themselves for precision and do not rely on the (sometimes cryptic) summaries provided by the search engines. The searches were carried out six times during a period of two months.

The purpose of this paper is

1. to learn about the number of different URLs retrieved on the given query (without examining the relevance of the documents) by all the search engines and by each separately and to learn about changes in the number of retrieved documents over the period of time the searches were carried out;
2. to examine the *overlap* among the results of the different search engines, where overlap between two engines is defined as the number of different URLs retrieved by both of the search engines on the given query;
3. to calculate the precision of the search process as a whole and the precision of each search engine on the given query, and as a side-effect to study the search engines' claimed ability to differentiate between proper names and nouns and their ability or inability to deal with non-English language documents;
4. to give an estimate on the recall of the whole process, and on the recall of each search engine separately.

2. Methodology

In order to explore the purpose of this paper, the search term used was the proper name Erdos, the name of the famous mathematician who passed away in September, 1996. Such search naturally returns references to other persons by the name of Erdos, and to places like the Erdos Cafe in Budapest, but should also contain all pages concerning the mathematician. Using the capital "E" in the search is supposed to return only pages where the name Erdos appears with capital "E" (i.e., a proper name). In Lycos we had to add a period after the name Erdos in order to get an exact match, otherwise Lycos automatically used word completion and returned also results with names like Erdossy, Erdosvary etc. (see (Lycos, 1996), this option is not existent or necessary in the version of Lycos as of December 1997).

The queried search engines (in alphabetical order) were:

- AltaVista (http://www.altavista.digital.com/cgi-bin/query?pg=aq),
- Excite (http://www.excite.com),
- Infoseek (http://www.infoseek.com),
- Lycos (http://www.lycos.com),
- Magellan (http://www.mckinley.com),
- Opentext (http://index.opentext.net/main/powersearch.html),
- Yahoo (http://www.yahoo.com).

Yahoo is not a search engine, but a directory service, containing documents classified by humans. If it does not find any relevant documents in its catalogue, it automatically passes the query on to AltaVista. This was the case with our query, no relevant documents were found by Yahoo. Therefore we compare the results of the remaining six search engines: AltaVista, Excite, Infoseek, Lycos, Magellan and Opentext.

The searches were carried out six times during a two months period between 21st November, 1996 and 27th January, 1997. A search was carried out on the 21st of each month (November 1996, December 1996 and January 1997). On these dates we searched all of the above mentioned search engines, and saved the search results. We repeated these searches immediately after we finished collecting the relevant documents resulting from the searches. These additional searches were carried out on the following dates: 30th of November, 1996; on the 28th of December, 1996 and on the 27th of January, 1997. Each of these six dates is called a search round.

It took us about a week to ten days to analyze the results of each search round and to collect the relevant documents from the Web. First, the query results were saved. A search engine displays ten to forty links to relevant documents on one page. Since hundreds of relevant documents were found, we had to save tens of pages of search results for each search engine. Next the URLs (the addresses of the Web documents, for a definition of URL, see for example (*Krol*, 1992)) the titles and other relevant information on each document (like date, size and relevancy grade given by the search engine, but not the summary) were filtered out from the pages returned by the search engines using a Visual Basic program. The display of the results and the additional information given varies from search engine to search engine, and sometimes even changed between the dates the searches were carried out. The output of this filtering process was a table with columns for the URL, the title and a column for each type of additional information. These tables were loaded into Microsoft Excel. On each table, corresponding to the results of a specific search engine on a specific date, we ran a Visual Basic module in Excel in order to identify *duplicates* – the same URL returned as the result of the given query more than once by the search engine at the current search round. Next, we compared the list of the currently returned URLs by the given search engine, to the list of URLs retrieved by the same search engine on the previous searches, again by running a Visual Basic module in Microsoft Excel. Then we constructed the list of all the new URLs retrieved by all the search engines in a given search round and we checked for URLs returned by more than one search engine. The list of nonduplicate URLs found in the current round was compared to the complete list of the URLs from the previous search round. The meaning of a URL appearing in both lists, is that the given URL was found by one search engine in a previous round, but was only found by another search engine in this round. This phase was also carried out by running a Visual Basic module in Microsoft Excel. The output of this phase was an Excel sheet containing the list of URLs that were first found in the current search round, called the list of *new URLs*. From this table, using a Visual Basic program, we created html pages with links to these new URLs with 50 links on each page. In the last phase of each search round each of these links were retrieved and the documents were saved on our local harddisk. This phase was carried out by "brute force", i.e., by connecting to the Web on about ten computers concurrently from our computer lab, and saving documents in parallel on each of them. The number of retrieved documents for a search round varied between 2926 and 482.

The analysis of the results was carried out by constructing frequency and cross tables and by utilizing the filtering tool of Microsoft Excel. The retrieved documents were manually categorized using the methods of content analysis (for example, see (*Krippendorff*, 1980)). The categories used appear in the subsection 3.4.

Using a single word, Erdos, in our searches, eliminated the differences between the search engines that result from their different retrieval capabilities on more complex searches, and also enabled us to give a clearcut definition of a precise document. The only difficulty with the chosen word, is that Paul Erdos was Hungarian, and thus his name is actually spelled Erdős (with a diacritic on the o). AltaVista claims to retrieve as a result of the query "Erdos" also documents that contain Erdos with diacritic on the o (without any mention of "Erdos" without diacritics in the document), while other search engines do not define in their help sheets whether they retrieve such pages, and conceivably ignore such documents. In the paper, we discuss the diacritic's influence on the results.

Note that:

1. in the context of this paper, identical documents are considered different if their URLs are different;
2. the comparison between the URLs is case sensitive i.e., http://207.63.103.38/Pub/math.html and http://207.63.103.38/Pub/Math.html are considered two different URLs. This is the policy of most of the search engines, since most of the servers are case sensitive;
3. if a link to a URL was found at a search round, and then again at one of following rounds, the corresponding document was retrieved only the first time the URL was found, thus we cannot say anything about the updates and changes that might have occurred to the documents during this period;
4. some search engines limit the number of displayed links that are relevant to the query, even though they state that they found a larger number of matching documents.

3. Results and discussion

3.1. Total retrievals

During the six search rounds, 6681 different URLs were retrieved as the results of the query "Erdos". Table 1 displays the number of different URLs retrieved by search round and by search engine. For some, unknown reason, sometimes Excite and Infoseek displayed exactly the same URL in a given search round more than once, for these two engines we placed the number of URLs retrieved by these search engines – including duplicates – in parentheses. Infoseek displays only the first 500 URLs, but states the actual number of URLs – these numbers are placed in brackets. AltaVista displays at most 1000 URLs for a given query, but has the additional ability to retrieve URLs that were created (or indexed by AltaVista) between given dates. This feature enabled us to partition the search results into four disjoint sets, each of size less than 1000. We became aware of this ability only from the third search round on. For the first two rounds, the number of URLs found is placed in brackets. The last row in the table, the total number of distinct URLs found in a search round, is calculated from the URLs actually retrieved, i.e. from the 500 displayed for Infoseek and from the 1000 for AltaVista in the first two rounds. In Table 1 we display the total number of distinct URLs per search round, as well as the total number of *new URLs* (URLs that were first found in the current round) per search round.

Note that the total number of relevant documents increases from search round to search round. Magellan and Opentext did not perform very well in our study, they were completely static during the two months period, they returned exactly the same links in each search round, even though Opentext states, that "The Open text index is updated continuously. We add and update over 50,000 Web pages per day" (Opentext, 1996a). All the other engines, except Excite (and a very slight decrease in AltaVista for the third round), showed a steady increase in the number of relevant documents found. Also note that URLs disappear between search rounds (a well-known phenomenon on the Web). If this was not the case, we would expect the total number of distinct URLs for Round 1, for example, to be 2926+482=3408, but the actual number is only 3023.

In Table 2, a different aspect of the total retrievals is examined. It displays the number of new URLs for each search engine in each search round. In the table we abbreviated "round" as "R". Note that the sum of the percentages in the last column exceeds 100, since some of the URLs were retrieved by more than one search engine.

Table 1
The number of URLs found in each search round broken down by search engine

Search engine	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
AltaVista	1000	1000	1937	2168	2169	2169
	[1969]	[1969]				
Excite	1494	1442	1165	1165	1376	1620
	(1982)	(1872)	(1647)	(1647)	(1788)	(2006)
Infoseek	468	468	497	497	500	500
	(500)	(500)	(500)	(500)	(500)	(500)
	[1007]	[1020]	[1375]	[1375]	[1378]	[1387]
Lycos	397	397	484	484	533	533
Magellan	139	139	139	139	139	139
Opentext	231	231	231	231	231	231
Total distinct URLs per round	2926	3023	3360	3891	3995	4220
Total distinct new URLs per round	2926	482	1203	813	625	633

Table 2
The number of new URLs by search engine and search round and the total number and percentage of distinct URLs by search engine

Search engine	R. 1	R. 2	R. 3	R. 4	R. 5	R. 6	Total number of distinct URLs per search engine	% of distinct URLs per search engine out of the total number of distinct URLs (6681)
AltaVista	1000	1	949	1090	4	0	3044	45.56
Excite	1494	488	423	0	836	788	4029	60.31
Infoseek	468	8	360	0	4	3	843	12.62
Lycos	397	128	6	0	96	0	627	9.38
Magellan	139	0	0	0	0	0	139	2.08
Opentext	231	0	0	0	0	0	231	3.46

The search engines have two means of discovering Web documents: by their robots that "crawl" in the Web and collect information on the URLs they visit; and by adding to their database URLs submitted to them by individuals. The first method is the main source for discovering Web pages. The numbers in Table 2 seem to indicate that, in the beginning of 1997, most search engines did not update their databases continually with the information retrieved by the robots (as they claimed), but rebuilt their indices periodically, and in between updated the indices by adding, probably, only self-submitted URLs.

The small number of URLs being discovered by Infoseek might partially be due to the fact that in each search round only 500 results were retrieved (out of 1007-1387 found by Infoseek).

These results are rather disappointing: none of the search engines cover the retrieved URLs very well (the values range between 2% and 60%!). We can come up with three possible explanations for this situation:

1. The set of URLs in the search engines' databases differ greatly, i.e., the crawler of each search engine crawls elsewhere and they cover largely different areas of the Web, in spite of the search engines' claim that each covers the whole Web: "Excite searches the entire Web for documents..." (Excite, 1997), or "Using AltaVista, you can find any word in any document published on the World Wide Web...Want to find all pages on the Web that contain information about Mars? AltaVista is the place." (AltaVista, 1997a) or "The experience with these earlier projects helped the researchers understand, despite conventional wisdom to the contrary, that indexing the entire Web might be feasible." (AltaVista, 1997b).
2. Each search engine interprets the search terms differently. This is highly unlikely in our case, where a single search term was used. Slight differences do occur due to diacritics and to the fact that some of the search engines are case insensitive (i.e. search engines also retrieve documents in which the first letter in the string "erdos" is not capitalized).
3. The crawlers of the different search engines visit the same URL at different times. When the robot of search engine X visits the URL, the document is relevant, but when the robot of search engine Y visits the same URL, the contents has changed and the document is not relevant anymore (or vice versa).

The examination of the documents suggests that the first point serves as a major explanation, thus it seems that in spite of their claims, none of the search engines is even close to indexing the whole Web. This is only a limited criticism, since indexing the whole Web might very well be an impossible mission at the current rate of growth and change of the Web. The search engines, however, should be more careful about their statements to avoid creating the wrong impression.

3.2. The difficulty with the diacritic

As we stated in the Introduction, the name of the mathematician, Erdos, is written with two dots or with two straight lines over the o in Hungarian (an o with an *umlaut*). Sometimes his name is spelt this way also in documents written in other languages.

AltaVista was the only search engine among the ones we examined, which stated that it retrieves both documents in which the search words appear without diacritics and documents in which the search words appear with diacritics (AltaVista, 1996). In the case of the letter o, the diacritics can be ò, ó, ô, õ, ö – as defined in the HTML coded character set by the W3 organization (W3, 1997). Infoseek stated that it gives "search support for European character sets" (Infoseek, 1996a) but probably only if words with accents are entered into the search field. The other search engines did not define how they act in this case. We consider a Web page an *umlaut document* if the name Erdos appears in it *only* with an umlaut. Documents where the name Erdos appears both with and without an umlaut are not considered umlaut documents. On the other hand, if in a document, the word "erdos" (with a small e, not capital e) appears without an umlaut, but Erdos (with a capital E) appears only with an umlaut, the document is considered an umlaut document, since when looking for a proper name, only words that start with a capital letter should be considered. Altogether 677 umlaut documents were found, which is 10.1% of the total number of documents. The breakdown by search engine is presented in Table 3.

Table 3
Umlaut documents by search engine, numbers and percentages out of the total number of documents retrieved by the search engine

Search engine	# of umlaut documents	% of umlaut documents out of the documents retrieved by the search engine
AltaVista	630	20.7
Excite	37	0.9
Infoseek	8	0.9
Lycos	6	1.0
Magellan	0	0
Opentext	13	5.6

It can be clearly seen that except for AltaVista, the other search engines retrieve umlaut documents only by accident. When the 63 umlaut documents not retrieved by AltaVista were examined more carefully, it turned out that in all but 4 of them the word "erdos" or "erdoshp.htm" or "erdosprog.html" appeared (without an umlaut and with small "e"). This probably explains why these documents were retrieved, although when looking for Erdos, documents containing this string with a little "e" should not be considered, and documents in which additional characters appear after the string "erdos", should definitely not be retrieved.

The search engines' disregard for the European character set is quite problematic. We tried to use the search terms "Erd&oumls" or "Erd&otildes" (sometimes the name is spelled Erdös with two dots on the o, and sometimes Erdős – with two lines on the o), where &ouml is the html encoding of ö and &otildes is the encoding of ő, in order to retrieve umlaut documents, but the query was not understood by the search engines. It is possible to enter the search term Erdös using the European character set, but there is no straightforward method to do this using the English-Hebrew character set (except for an extremely tedious procedure involving cut-and-paste, that we came upon by chance). The number of non-English documents rapidly increases along with the number of English language documents, thus hopefully the search engines will soon address this problem. AltaVista's solution is not ideal either, since there is no option in AltaVista to retrieve only non-umlaut documents.

The simplest general solution is to start using unicode (a 16-bit character set, large enough to accommodate all the languages in the world with no overlapping codes) instead of the different versions of ASCII currently used (8-bit character sets with the same code for different characters in different languages). For more information on unicode, see for example (*McClure & Stan*, 1995) or The Unicode Homepage on the Web (1997). In a recent article *Oudet* (1997) discussed multilingualism on the Internet.

3.3. Overlap

The *overlap* between two search engines is the set of URLs retrieved by both of the search engines on our query during the six search rounds, i.e., the intersection ($\cap$) of the sets of URLs retrieved by each search engine. If a URL was retrieved by one search engine in a search round, and was returned by the second engine in a different search round, it is still considered an element of the overlap set. This definition of overlap can naturally be extended to more than two search engines. The sizes of these sets are displayed in Table 4. The names of the search engines are abbreviated by the first letter of each search engine's name. For the sake of completeness, the sizes of the sets of URLs retrieved by each engine separately are also displayed. Note that the intersection sets are disjoint, i.e., "$A \cap E$ exactly" is the set of URLs that were found by AltaVista and Excite and by none of the other engines, more formally "$A \cap E$ exactly" is the set $A \cap E \cap \bar{I} \cap \bar{L} \cap \bar{M} \cap \bar{O}$, where $\bar{X}$ is the complement of the set X. Similarly, "E exactly" is the set of URLs that were found by Excite and by none of the other search engines, formally "E exactly" is the set $E \cap \bar{A} \cap \bar{I} \cap \bar{L} \cap \bar{M} \cap \bar{O}$.

Table 4
The number of URLs in each of the disjoint overlap sets

Sets	# of URLs	Sets	# of URLs	Sets	# of URLs
A exactly	1680	$A \cap E \cap I$ exactly	158	$A \cap E \cap I \cap L$ exactly	59
E exactly	2622	$A \cap E \cap L$ exactly	74	$A \cap E \cap I \cap M$ exactly	2
I exactly	315	$A \cap E \cap M$ exactly	15	$A \cap E \cap I \cap O$ exactly	13
L exactly	275	$A \cap E \cap O$ exactly	37	$A \cap E \cap L \cap M$ exactly	4
M exactly	57	$A \cap I \cap L$ exactly	18	$A \cap E \cap L \cap O$ exactly	4
O exactly	58	$A \cap I \cap M$ exactly	4	$A \cap E \cap M \cap O$ exactly	2
$A \cap E$ exactly	788	$A \cap I \cap O$ exactly	7	$A \cap I \cap L \cap M$ exactly	3
$A \cap I$ exactly	78	$A \cap L \cap M$ exactly	3	$A \cap I \cap L \cap O$ exactly	0
$A \cap L$ exactly	32	$A \cap L \cap O$ exactly	1	$A \cap I \cap M \cap O$ exactly	1
$A \cap M$ exactly	17	$A \cap M \cap O$ exactly	3	$A \cap L \cap M \cap O$ exactly	0
$A \cap O$ exactly	34	$E \cap I \cap L$ exactly	15	$E \cap I \cap L \cap M$ exactly	1
$E \cap I$ exactly	107	$E \cap I \cap M$ exactly	1	$E \cap I \cap L \cap O$ exactly	1
$E \cap L$ exactly	58	$E \cap I \cap O$ exactly	3	$E \cap I \cap M \cap O$ exactly	0
$E \cap M$ exactly	6	$E \cap L \cap M$ exactly	2	$E \cap L \cap M \cap O$ exactly	0
$E \cap O$ exactly	40	$E \cap L \cap O$ exactly	10	$I \cap L \cap M \cap O$ exactly	0
$I \cap L$ exactly	45	$E \cap M \cap O$ exactly	2	$A \cap E \cap I \cap L \cap M$ exactly	0
$I \cap M$ exactly	2	$I \cap L \cap M$ exactly	1	$A \cap E \cap I \cap L \cap O$ exactly	4
$I \cap O$ exactly	1	$I \cap L \cap O$ exactly	1	$A \cap E \cap I \cap M \cap O$ exactly	0
$L \cap M$ exactly	8	$I \cap M \cap O$ exactly	0	$A \cap E \cap L \cap M \cap O$ exactly	0
$L \cap O$ exactly	4	$L \cap M \cap O$ exactly	1	$A \cap I \cap L \cap M \cap O$ exactly	2
$M \cap O$ exactly	1			$E \cap I \cap L \cap M \cap O$ exactly	0
				$A \cap E \cap I \cap L \cap M \cap O$ exactly1	

The results appearing in the above table are quite surprising. To emphasize the results, we summarized them in Table 5. In this table we partitioned the set of URLs by the number of search engines that found that URL. Documents belonging to the category "exactly one" were found during the search rounds by one search engine only, this search engine could be any one of the search engines that were used.

Table 5
Breakdown of the retrieved documents by the number of search engines that retrieved them, absolute numbers and percentages

Number of search engines	Number of URLs	Percentage of URLs out of the total number of retrieved URLs (6681)
exactly one	5007	74.94
exactly two	1221	18.27
exactly three	356	5.33
exactly four	90	1.35
exactly five	6	0.09
exactly six	1	0.02

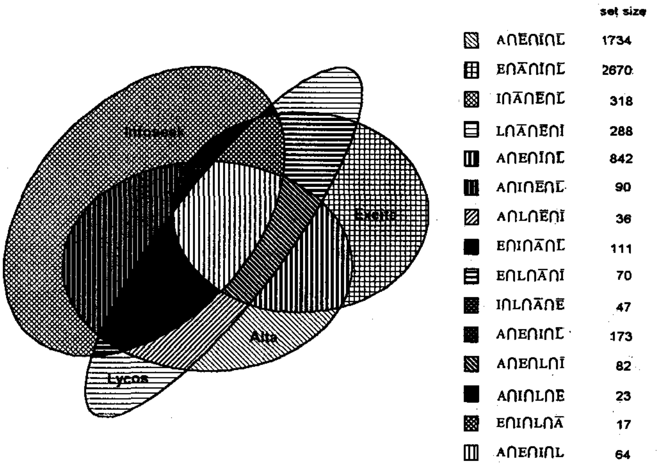

Fig. 1. Overlap

It turns out that nearly 75% of the URLs were found by one search engine only, and only a single document (0.02% of the documents) was retrieved by all six participating search engines during the search rounds.

The single document retrieved by all six search engines is entitled "POP Mathematics" (its URL is http://archives.math.utk.edu/popmath.html). Among other popular mathematical subjects, it describes and links to the "Erdos number project". The name Erdos appears seven times in the middle of the document. All engines, except AltaVista found this page already in the first round. Excite ranked it 51% (the highest relevance in this round was 78%), Infoseek 50% (the highest relevance in this round was 56%), Lycos 87% (the highest relevance was 100%).On the ranked lists of Magellan and Opentext, the document appeared in the first (out of 139) and the 42nd (out of 231) places respectively. AltaVista's advanced search is unranked. In AltaVista's list, this URL first appears in round 4 (28th December, 1996), even though AltaVista dates the document to the 6th December, 1996 (for an undated document like "POP Mathematics", this is the date when AltaVista first discovered this URL), and the previous search round was carried out on the 21st December, 1996, thus could have appeared in the search results already in round 3.

Even if we disregard the results of the two "static" engines, Magellan and Opentext, and consider only the four remaining ones, the results do not change dramatically as can be seen from Fig. 1 and Table 6. In this case the total number of distinct URLs retrieved by the four search engines during the search rounds is 6565. The only significant difference between Tables 5 and 6 is in the number of URLs retrieved by all of the search engines (1 versus 64).

Table 6
Breakdown of the retrieved documents by the number of search engines that retrieved them, absolute numbers and percentages for the four 'large' search engines

Number of search engines	Number of URLs	Percentage of URLs out of the total number of retrieved URLs (6565)
exactly one	5010	76.31
exactly two	1196	18.22
exactly three	295	4.49
exactly four	64	0.98

To strengthen the impact of these results, we bring as comparison the results of another search. This search was carried out on the 3rd April 1997. The query was formulated as *bibliometrics AND growth* (or for search engines not capable of Boolean search the query was formulated as: *+bibliometrics +growth).* The participating search engines were AltaVista, Excite, Hotbot (http://www.hotbot.com), Infoseek, Lycos and Opentext. In this case we collected the URLs the search engines pointed to without actually retrieving the documents themselves. Altogether 146 distinct URL were found according to the breakdown appearing in Table 7. Note that the sum of the percentages is greater than 100, because of the overlaps. In Table 8, we calculated how many search engines found each URL.

Table 7
Numbers and percentages of distinct URLs per search engines for the query: bibliometrics AND growth

Search engine	# of distinct URLs	% of distinct URLs out of the total number of URLs (146)
AltaVista	64	43.8
Excite	37	25.6
Hotbot	82	56.2
Infoseek	1	0.7
Lycos	0	0
Opentext	6	6

Table 8
Breakdown of the retrieved documents by the number of search engines that retrieved them, absolute numbers and percentages for the four 'large' search engines for the query: bibliometrics AND growth

Number of search engines	Number of URLs	Percentage of URLs out of the total number of retrieved URLs (146)
exactly one	109	74.66
exactly two	28	19.18
exactly three	8	5.48
exactly four	1	0.68
exactly five	0	0
exactly six	0	0

The similarities in the percentages in Tables 6 and 8 are striking, especially for the URLs retrieved by a single search engine only. Thus it seems that one cannot dismiss these results as a coincidence. The breakdown by search engine is presented in Table 9.

Table 9
Breakdown of the numbers and percentages of the URLs retrieved by search engine and by the number of other search engine that retrieved it.

Search engines	# of URLs	% of URLs out of total for search engines	Search engine	# of URLs	% of URLs out of total for search engines
AltaVista only	1734	57.0	Excite only	2670	66.3
AltaVista + 1 other engine	968	31.8	Excite + 1 other engine	1023	25.4
AltaVista + 2 other engines	278	9.1	Excite + 2 other engines	272	6.7
AltaVista + 3 other engines	64	2.1	Excite + 3 other engines	64	1.6
Infoseek only	318	37.7	Lycos only	288	45.9
Infoseek + 1 other engine	248	29.4	Lycos + 1 other engine	153	24.4
Infoseek + 2 other engines	213	25.3	Lycos + 2 other engines	122	19.5
Infoseek + 3 other engines	64	7.6	Lycos + 3 other engines	64	10.2

It might have been the case that there is one outstanding search engine (at least for our query) and all the others performed very badly. This claim is ruled out by the results presented (for the original query "Erdos") in Table 9. Again we disregard the results of Magellan and Opentext. From the table, it can be seen that the behavior of AltaVista and Excite is similar, so is the behavior of Infoseek and Lycos. The most outstanding result is that 66.3% of the documents found by Excite were retrieved by this search engine alone, and for AltaVista the value of the same variable is 57%.

3.4. Precision

Before jumping into conclusions, one should check the precision of the retrieved documents. In the context of the current paper, a document is considered *precise* if it contains the word Erdos (with or without an umlaut) with a capital "E". Note hat previous papers ((*Leighton & Srivastava*, 1997) and (*Ding & Marchionini*, 1996)) discuss the problems in deciding what is relevant and precise for complex queries. In our case the query is simple and it is straightforward to give a clear, operational definition of precision.

The retrieved documents were partitioned into the following categories:

Erdos – the precise documents (to enable further research, this category was further subdivided into documents about the mathematician Erdos and other persons or places named Erdos);

erdos – documents in which the string "erdos" appeared with little "e" only;

erdos in URL – the string erdos (usually with little e) appears in the URL only;

no erdos – documents in which the string "erdos" (with or without capital "E" and with or without an umlaut) does not appear at all;

not existent – when trying to connect to the URLs retrieved by the search engine, a "NOT FOUND" message was received by the browser;

inaccessible – documents that could not be reached because of problems with the communication network or with the servers on which they reside. Several attempts were made over a period of time to reach these URLs.

Documents belonging to the categories "erdos" and "erdos in URL" probably result from the fact that some of the search engines seem to be case insensitive. Indeed, Magellan and Opentext state in their help sheets that they are case insensitive: "It doesn't matter whether you use capital or lowercase letters" (Magellan, 1996), and "Open Text ignores case" (Opentext, 1996b). AltaVista and Infoseek both claim to be case sensitive, and in fact, have not retrieved any documents in which the string "erdos" occurred only with a small e. Lycos does not say anything clear about case sensitivity. The greatest disappointment in this area was Excite, which tells the user: "Searching for Proper Names...just capitalize the first letters of each word" (Excite, 1996), but retrieved the largest number of documents in the categories "erdos" and "erdos in URL".

A document can belong to the categories "no erdos" or "not existent" for one of three reasons:

1. The document changed or disappeared between the time it was indexed and the time it was retrieved.
2. The document changed or disappeared between the time we saved the link and the time we actually retrieved the document.
3. The document was not indexed correctly.

The third point is not very likely due to the fact that the overall precision of the searches is quite good. Reasons 1 and 2 have to be taken into account. We made every effort to lessen the effect of the second point by trying to retrieve the documents as fast as possible. It took us about a day to construct the list of the new URLs after which we immediately started to collect the documents (almost 3000 documents in the first search round). The document collection itself took about a week, and it is quite conceivable that some of the documents changed or disappeared during this period. Careful analysis per search round of the number of new documents in the categories "no erdos" and "not existent" does not show any clear pattern in favor of the second point over the first one. Most probably the search engines do not (and cannot) update their indices often enough to eliminate the problems that arise from changing or disappearing documents. Thus the category "not existent" is here to stay.

The category "inaccessible" is due to communication problems and to servers that were down every time we tried to access the URLs. Altogether 134 (2.0% of the total number of URLs), documents could not be retrieved. These URLs should not be taken into account when computing the precision of our search process, and only the set of accessible URLs should be considered, called the set of *accessible documents*. The size of this set is 6681-134=6547.

Thus we define:

$$precision = \frac{number\ of\ precise\ documents}{number\ of\ accessible\ documents} * 100\%$$

Table 10
The numbers and percentages of documents in each category

Category	# of documents	% of total accessible (6547)
Erdos	5055	77.2
erdos	391	6.0
erdos in URL	63	1.0
no erdos	279	4.2
not existent	759	11.6

In Table 10, the number of documents in each category (except for inaccessible) is displayed and the percentages are calculated out of the total number of accessible documents.

The precision of the search process is: 77.2%. This number is unexpectedly high considering remarks in previous works (e.g. *Feldman* (1997b): "every search engine will give you surprisingly bad results some of the time" and "no search engine finds it all"). *Venditto* (1996) pointed out that, "each one delivered a high proportion of irrelevant information when challenged with anything beyond a simple search on a well-represented topic". Our query was extremely simple, which could explain the high precision of the retrievals.

The breakdown into categories of each search engine separately is displayed in Table 11. Again the percentages are calculated out of the number of accessible documents for each search engine. The number in parentheses under the name of the search engine is the number of accessible documents for that search engine.

Table 11
The numbers and percentages of documents in each category per search engine

	Alta (3009)		Excite (3940)		Info (843)		Lycos (623)		Mag. (139)		Open (231)	
Cat.	#	%	#	%	#	%	#	%	#	%	#	%
Erdos	2568	85.3	3137	79.6	704	84.4	468	75.1	81	59.2	174	75.3
erdos	0	0	351	8.9	0	0	31	5.0	1	0.7	12	5.2
erdos in URL	0	0	61	1.5	0	0	4	0.6	0	0	5	2.2
no erdos	119	4.0	77	2.0	42	5.0	43	6.9	27	19.7	24	10.8
not existent	322	10.7	314	8.0	88	10.6	77	12.4	28	20.4	16	6.9

The search engine with the highest precision is AltaVista with Infoseek following closely. Third place belongs to Excite. As we pointed out before, in spite of what appears in Excite's help sheets, it has a problem with case sensitivity. This explains why

10.4% (8.9% in "erdos" and 1.5% "erdos in URL") of the documents are irrelevant. Opentext, in spite of the fact that the search results were exactly the same in all of the six rounds, achieved quite a good mark in precision, and did slightly better than Lycos. Magellan is the least precise, covered the least number of documents and its search results have not changed at all during the search rounds.

3.5. Recall

During our search process we found 5505 relevant documents. This number can serve as an approximation for the total number of URLs in the Web at the beginning of 1997. We believe that this number is an under-estimate, but we cannot propose a better one. Using this number we can calculate the *approximate-recall* for each search engine, where approximate recall of a search engine is defined as:

$$\textit{approximate - recall} = \frac{\textit{total number of relevant documents retrieved by the search engine}}{\textit{estimated total number of relevant URLs}} * 100\%$$

The approximate-recall for each search engine appears in Table 12. To allow comparison, the percentage of distinct URLs per search engine out of the total number of distinct URLs retrieved during the search process is also displayed (these data appear in a more detailed format in Table 2).

Table 12
The approximate-recall and the fraction of distinct URLs of the search engines in percentages

Search engine	Approximate recall	Distinct URLs per search engine out of the total number of distinct URLs (6681)
AltaVista	46.6%	45.56%
Excite	57.0%	60.31%
Infoseek	12.8%	12.62%
Lycos	8.5%	9.38%
Magellan	1.5%	2.08%
Opentext	3.2%	3.46%

Note that only for AltaVista and Infoseek the approximate-recall values are higher than the respective values for the fraction of URLs returned out of the total number of URLs retrieved. The reason for this is the high precision of these two search engines.

These low and disappointing values for approximate-recall are caused by the combination of the small fraction of the total documents retrieved by each search engine, the small overlap between the search results of the different engines and the relatively high value for the precision of each of the search engines.

The only search engine with an "excuse" is Infoseek, since it displays only the first 500 URLs relevant to the search. The statement: "Lycos claims to have indexed 91 percent of the Web with a spider technology" appears more than once in the literature (see for example (*Wired*, 1997)), but it is very far from being supported by our findings.

Note that we only used a very conservative estimate for the total number of relevant documents in the Web. We suspect that the real number (as of the beginning of 1997) was easily at least twice (or even three times) our estimate. If this is the case, the actual recalls are the values presented here divided by two (or three)!

4. Conclusion

The main finding is that there is no one "best search engine", at least when one is interested in high recall. A search engine may be better than another in terms of its user interface or search capabilities, but no search engine even comes close to covering the whole Web (at least for the queries "Erdos" and "bibliometrics AND growth"). As we said before, it may be impossible to cover the whole Web, because of its rapid rate of change and growth, and the search engines should make sure that their users are aware of this fact.

The high precision of the search engines came as a pleasant surprise. It would be interesting to find out why people assume that the search engines have low precision. One possible explanation, which should be further investigated is that the search engines do not do so well on more complex queries involving more than one search term.

The number of Web pages not written in English increases constantly. The search engines have already started to pay attention to them. For example, as of December 1997, AltaVista introduced the option of choosing the language of the retrieved Web documents from 25 languages. We hope that in the future, problems arising from using different, overlapping character sets will be solved.

It is interesting to compare the Web to the Library of Congress: the WWW was released by CERN in 1991 (developed by Tim Berners-Lee). The first graphic browser *Mosaic* was developed by Marc Andreessen and his team at NCSA, but the World Wide Web really started to take off only after the release of Netscape 1.0 in 1994 (the historical data is based on (*Cailliau*,1995)). The Web was only about three years old by the end of 1996, and already incorporated at least 80 million pages: in November 1996, Infoseek reported that "it has found over 80 million unique URLs" (Infoseek, 1996b), while at the beginning of 1996, the estimated number of Web pages was 18 million (*Courtois*,1996, p. 35). Thus in 1996 alone, the Web grew by 62 million pages – and this number could very well be hugely underestimated!

In 1996, the 196 years old Library of Congress had 105 million items (see (*Feather & Struges*, 1997)). Its rate of growth is 7,000 items per day, or approximately 2 million items per year. The Web grew at least 30 times faster than the Library of Congress in 1996! A major difference between the items in the 'Library of Congress and the documents on the Web is that the items already belongiñg to the Library of Congress' collection do not change or disappear, thus the catalogue only grows, whereas the Web is dynamic: documents get updated or are removed from the Web, and new ones are constantly added to it. In some sense, the search engines' work is much harder than that of the cataloguers at the Library of Congress.

One additional issue, which was already pointed out by *Rousseau* (1997), deserves attention. The search results are irreproducible, since these results are highly dependent on the time the searches were carried out. The same applies for the retrieved documents. Thus, it is the researchers' responsibility to save these results on their computer systems, and to take care of appropriate backup, and be ready to produce these results on demand. For the current research this means saving hundreds of megabytes of data forever. Even this is not enough. In several places we refer to the help pages of the search engines, as they appeared by the end of 1996. Since then they have totally changed and sometimes the current versions are totally irrelevant, but luckily we printed them out before the beginning of the search rounds. There are already a few initiatives to save "snapshots" of the Web, i.e., to archive its contents (see (*Feldman*, 1997a)). One of the most extensive ones ,Internet Archives, is discussed by its founder in (*Kahle*, 1997).

To conclude, we believe that more such extensive and thoroughly analyzed searches and retrievals should be carried out in order to learn more about the true nature of the Web and its capabilities.

I would like to thank *Bluma Peritz* for introducing me to the subject of bibliometrics, for her support and encouragement and for her very helpful remarks.

References

AltaVista. (1996). *Help for Advanced Query.* [Online]. Available: http://altavista.digital.com/cgi-bin/query?pg=ah [November 1996].

AltaVista. (1997a) *About AltaVista Search.* [Online]. Available: http://www.altavista.digital.com/av/content/about.htm [December 1997].

AltaVista. (1997b) *About AltaVista Search.* [Online]. Available: http://www.altavista.digital.com/av/content/about_our_story_2.htm [December 1997].

ALMIND, T. C. & INGWERSEN, P. (1997). Informetric Analyses on the World Wide Web: Methodological Approaches to 'Webometrics'. *Journal of Documentation, 53(4),* 404-426.

BAR-ILAN, J. (1997). The 'Mad Cow Disease', Usenet Newsgroups and Bibliometric Laws. *Scientometrics,* 39(1), 29-55.

CAILLIAU, R. (1995). *A Little History of the World Wide Web.* [Online]. Available: http://www.w3.org/History.html [December 1997].

CHU, H. & ROSENTHAL, M. (1996). Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology. *ASIS96.* [Online].Available: http://www.asis.org/annual-96/Electronic-Proceedings/chu.htm [December 1997].

COURTOIS, M. P. (May/June 1996) Cool Tools for Searching the Web – An Update. *Online,* 29-36.

DEZELAR-TIEDMAN, C. (1997). Known-Item Searching on the World Wide Web. *Internet Reference Services Quarterly,* 2(1), 5-14.

DING, W. & MARCHIONINI, G. (1996). A Comparative Study of Web Search Service Performance. *ASIS96.* [Online]. Available: http://www.glue.umd.edu/~weid/asis/fulltext.htm [December 1997].

DONG, X. & SU, L.T. (1997). Search Engines on the World Wide Web and Information Retrieval from the Internet: A Review and Evaluation. *Online & CDROM Review,* 21(2), 67-81.

Excite. (1996). *How to use Excite search.* [Online]. Available: http://www.excite.com/Info/searching.html?an [November 1996].

Excite. (1997). *What We Do.* [Online]. Available: http://corp.excite.com/Company/what.html [December 1997].

FEATHER, J. & STURGES, P. (Eds). (1997). *International Encyclopedia of Information and Library Science.* London: Rutledge, 1997, 263-265.

FELDMAN, S. (1997a). 'It Was Here a Minute Ago!': Archiving the Net. *Searcher,* 5(9), 52-. [Also Online]. Available: http:// www.infotoday.com/searcher/oct/story4.htm [December 1997].

FELDMAN, S. (1997b). 'Just the Answers, Please': Choosing a Web Search Service. *Searcher, 5(5),* 44-57. [Also Online]. Available: http:// www.infotoday.com/searcher/may/story3.htm [December 1997].

HASKIN, D. (1997). Power Search. *Internet World, 8(12),* 79-92.

Hypertext Markup Language – 2.0 – The HTML Coded Character Set. (1997). [Online]. Available: http://www.w3.org/MarkUp/html-spec/html-spec_13.html. [December 1997].

Infoseek. (1996a). *About Ultraseek.* [Online]. Available: http://guide.infoseek.com/Help?pg=AboutUltra.html&sv=N3 [November 1996].

Infoseek. (1996b). *Feature Comparison.* [Online]. Available: http://guide.infoseek.com/doc?pg=comparison.html&sv=N3 [November 1996].

KAHLE, B. (March 1997). Preserving the Internet. *Scientific American,* 82-83.

KRIPPENDORFF, K. (1980). *Content Analysis – An Introduction to Its Methodology,* Beverly Hills: Sage Publications, 1980.

KROL, E. (1992). *The Whole Internet Guide,* New York: O'Reilly, 1992.

Lycos. (1996). *Lycos Inc. Information.* [Online]. Available: http://www.lycos.com/help.html [November 1996].

LARSON, R. (1996). Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace. *ASIS96.* [Online]. Available: http://sherlock.berkeley.edu/asis96/asis96.html [December 1997].

LANCASTER, F. W. & FAYEN, E.G. (1973). *Information Retrieval On-Line,* Los Angeles: Wiley-Becker. chapter 6.

LEIGHTON, H. V. & SRIVASTAVA, J. (1997). *Precision among World Wide Web Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos.* [Online]. Available: http://www.winona.msus.edu/is-f/library-f/webind2/webind.html [December 1997].

Magellan. (1996). *Magellan's Frequently Asked Questions.* [Online]. Available: http://www.mckinley.com/feature.cgi?faq_bd [November 1996].

MCCLURE W. L. & STAN A. H. (1995). Communicating Globally: The Advent of Unicode. *Computers in Libraries, 15(5),* 19-24.

Opentext. (1996a). *The Open Text Index – Frequently Asked Questions.* [Online]. Available: http://index.opentext.net/main/help.htm [November 1996].

Opentext. (1996b). *The Open Text Index – Search Help.* [Online]. Available: http://index.opentext.net/main/help.htm [November 1996].

OUDET, B. (March 1997). Multilingualism on the Internet. *Scientific American,* 77-78.

ROUSSEAU, R. (1997). Sitations: an Exploratory Study. *Cybermetrics,* [Online], 1(1). Available: http://www.cindoc.es/cybermetrics/articles/v1i1p1.htm [November 1997].

TOMAIULO N. G. & PACKER, J. G. (1996). An Analysis of Internet Search Engines: Assessment of Over 200 Search Queries. *Computers in Libraries, 16(6),* 58-62.

The Unicode Homepage on the Web. (1997). [Online]. Available: http://www.unicode.org [December 1997].

VENDITTO, G. (1996). Search Engine Showdown. *Internet World, 7(5),* 79-86.

Wired Cybrarian. (1997). [Online]. Available: http://www.wired.com/cybrarian/frame/reference/stats.html [December 1997].

WOODRUFF, A. et al. (1996). An Investigation of Documents from the World Wide Web. *Proceedings of the Fifth International World Wide Web Conference.* 963-980.

ZORN, P. et al. (May/June 1996). Searching – Tricks of the Trade. *Online,* 15-28.