Search and Discovery on the Web

Bipin C. Desai

Department of Computer Science
Concordia University
1455 de Maissonneuve Blvd. West
Montreal, Canada H3G 1M8
bcdesai@cse.concordia.ca
http://users.encs.concordia.ca/∼bcdesai

Fall 2001

Abstract. In this paper we report on a series of tests we have done over three generations of search engines and comment on the completeness of their coverage. The existing and new search systems, while becoming more focused, tend nonetheless to generate misses and false hits; this is due to the fact that they attempt to match the specified search terms without context in the target information resource. Our tests, repeated over three generation of search engines, over the last few years show that while the recall for a simple search has increased, the precision has actually decreased. We give some reasons for this decrease and suggest the need for increasing the intelligence of the search process by adding semantics information to the indexing scheme.

1 The Search and Discovery Problem

The Internet has become extremely popular through the medium of the Web which is now just over ten years old. Over these ten years, its growth has been dramatic, helped in a large part by graphical browsers, one of the first widely used being Mosaic. It is estimated that each day, the number of Web pages increases by one million. It is recognized that the Internet, via the intermediary of the Web, has become a readily accessible repository of information and has become the accepted norm for disseminating and sharing information resources in hyper-media.

This wealth of information is interconnected by billions of hyper-links. Billions of people have access to this information through browsers. Being a decentralized and democratic system, there is no structure to the large amount of information accessible via the Web and the Internet. From the very beginning, a number of pioneers of the web had recognized the need for and development of a search and discovery system for the web [1–3]. A number of early search systems were introduced in the mid 1990's and their performance and co:operation were the subject of a WWW-III workshop[4]. The search engines have evolved over the last seven years and today most users, when searching for information about a specific topic, turn to one or more search engines.

In this paper we report on a series of tests we have done, over what we call the three generations of search engines, and examine the completeness of their coverage. In the following section we briefly describe the anatomy of search engines. In section 3, we give the methodology used in evaluating the search engines and in section 4, the results of our tests. Conclusions are reported in the final section of this paper.

2 Anatomy of a Search Engine

Rapid growth in data volume, user base and data diversity render Internet-accessible information increasingly difficult to use effectively. With this mass of Internet information, efficient indexing and retrieval of electronic information resources become more critical to an expanding population of users. The existing and new search systems, while becoming more focused, tend nonetheless to generate misses and false hits; this is due to the fact that they attempt to match the specified search terms without context in the target information resource.

Information Retrieval (IR) is concerned with the representation, storage, organization and accessing of information. The first step in the retrieval process is for the user to state the information needed. This has to be done in a format that enables the IR system to understand it and to act on it[5]. To facilitate the task of finding items of interest, libraries and information centers provide information users with a variety of auxiliary aids. Each incoming item is analysed and appropriate descriptions are chosen to reflect the information content of the item. Retrieval effectiveness is typically measured by two metrics: precision, which is the percentage of the retrieved documents that are relevant to the information need, and recall, which is the percentage of relevant documents in the collection that are retrieved[5]. Indexing is the basis for retrieving documents that are relevant to the user's need[6]. Compact descriptions of a document's index may increase the efficiency of matching and the effectiveness of classifying textual material as relevant or non-relevant. Document retrieval imposes conflicting normalising and accuracy demands[6]. As a result, variations in indexing that increase precision usually decrease recall, and vice versa. The fundamental goal is to increase both.

In conventional information retrieval, the stored records are normally identified by sets of keywords or phrases known as index terms. Requests for information are typically expressed by boolean combinations of index terms. The retrieval system is designed to select those stored items that are identified by the exact combination of search terms specified in the queries. The terms characterising the stored texts may be assigned manually by trained personnel, or automatic indexing methods may be used to handle the term assignment.

However, with the amount of information on the Internet, manual indexing becomes very expensive. Even though there are a number of systems which rely on either volunteer or paid editors to index Internet resources, most search engines use programmed systems to both locate information resources and index them.

There are a number of excellent articles which describe the anatomy of search engines[7, 8]. The components of a typical search engine are a web-crawler which collects information resources from the web, a document processor, a Web based user interface for entering the search request and a query processor to handle such requests; a search subsystem, a ranking and categorizing subsystem and finally a subsystem to present the results to the users in blocks of some predetermined sizes. In processing the document, the critical component is to identify the terms or indexable components of the document and add these to an inverted index. In this process the usual steps consists of deleting stop words, stemming terms to a common root form, assigning weight to terms which is based on a combination of the frequency and position of their occurrence, and the ratio of text to inverse document frequency(TF/IDF).

It has been recognized by research librarians that good query gives good retrieval. Unlike the human system, the Web based query interface has limitations and most searches involve conjunction and/or disjunction of terms. The facilities provided in the early search engines have been expanded to allow advanced query interface to include strings, adjacency or proximity operations. However, most of these systems still do not allow search to specify the context of terms used in the search.

The terms in the query are first stemmed to the the root format before they are used in the search process. Some search engines assign weights to the terms in the user query based on the order of the terms. Other pre-processing steps may be used before the search for the user supplied terms in the inverted indices is attempted. The resulting set of web pages are usually ranked and the ranked list is presented to the user. The ranking is done so as to present the most relevant documents at the top of the lists. Different search engines use different models in determining the ranking. The third generation search engines go a step further and categorize the results and present the user with both the categorized sets as well as the raw ranked lists.

Since search engine companies have not been successful in charging for the search service, their revenue is derived from either publicity on the pages presented to the user or by selling the top entries in the search results to commercial sponsors. This latter makes the user suspect of the results of the search and this practice is being debated[9].

3 Searching and Discovery using Search Engines

In IR, a number of measures are used to express the effectiveness of an information search operation. The first one is recall and the second is precision. Relevance of a document is determined by the users and is based on whether it responds to the information needs. Recall is the proportion of relevant documents returned by the search system and precision is the proportion of the documents actually relevant from the set that is found by the retrieval system[10]. We use these measures in our tests. Equation 1, gives R, the relevance of the recall or simply the recall and equation 2 gives R_a, the average recall over a number of search engines for the same search. Equation 3, gives P, the precision of the search and

equation 4 gives P_a, the average precision over a number of search engines for the same search.

$$R = \frac{\#\ relevant\ retrieved}{\#\ existing\ relevant} \quad (1) \qquad P = \frac{\#\ relevant\ retrieved}{\#\ retieved} \quad (3)$$

$$R_a = \frac{\sum_{i=1}^{n} R_i}{n} \quad (2) \qquad P_a = \frac{\sum_{i=1}^{n} P_i}{N} \quad (4)$$

In IR, the query operation is classified as being for the search for a single document or part of document, the selection of most relevant documents or all documents matching some criterion. With the search engine, the typical goal is to find relevant documents that contain a given set of words or phrases. Hence they do not usually support search based on most common retrieval tasks which consist of locating documents with a given title, author, or subject[11]. With the web, people have learned to use the name of the author, possible string in the title and likely subject or keywords used as terms in the search with the hope that the search engine would have indexed the relevant documents with these terms. Most of the times it seems to work and the user finds useful results but not necessarily all of it or the most relevant.

3.1 Search and Evaluation Methodology

Relevance of documents retrieved depends on the needs of the users. User may not always be aware of what exists in the database of the search engines nor what database it uses. Since the decision whether a URL produced by a search engine is relevant or not for a search is a task that is done by the searcher, it is a subjective decision. The usefulness of a document is related to the needs of the user and two different users with different needs may use the same search terms and may get very similar results for the search from the search engines; it is unlikely that both of them will use the same set of documents from the results of the search.

Another problem with search engines is their inconsistency. The same search when repeated need not produce the same results. The reason for this is that most search engines have been designed to provide a few relevant answers very quickly(in a few hundred milliseconds) and search consistency is sacrificed. Hence the results are not very accurate. Many a times even basic boolean processing may not work as expected. The problem most users face is not knowing what actually is on the Web when they start a search and hence they cannot judge whether the results did locate all the relevant documents or not. Furthermore, most users do not go beyond the hits displayed on the first few pages of the results. Our tests consist of checking what the search engine has located and indexed from a set of documents known to exist. The decision of whether a document is relevant or not in our case is made non-subjective by using the relevance criterion that the document should be authored or clearly be related to the author. In our case, we knew the exact number of documents accessible on

the Web. For our tests we define the terms hits, duplicates, mis-hits and misses as follows:

Hits: A URL is a hit if the corresponding document contains the search string and it pertains to the intent of the search, namely a document about or by the author.

Duplicates: If the same document is served from more a than one server then it is considered as a duplicate. Here the URLs are different but the contents are the same. This is one of the problems that has plagued search engines from the start; some search engines have addressed this problem better as our experiments illustrate.

Miss-hits: A URL is considered a miss-hit if the document is not relevant for the search. Here even though the search terms may exist in the document they occur out of context. Search engines again have difficulty with context of words in documents.

Misses: The number of relevant documents not found even though they existed on the Web. Since we started with a list of known URLs being served long before the tests, this was easy to determine.

In each of the tests we used the search terms "Bipin (AND) Desai"[1]. The boolean (AND) was used only when it was required by the search engine to represent a conjunction. In spite of the conjunction, some search engines returned pages which contained only one of the two search terms: this was especially true in the early days of the search engines. Even though we have designated the search engines to be in three generations, there has been a continuous development in the technology used by search engines and at times it is difficult to mark the generations. Lycos that was included in the first generation was one of the first major engines that was aggressive in its mission of crawling the web in its academic version. Lycos was commercialised and like the other commercial search engines, has evolved. The second generation was marked by introduction of search engines such as AltaVista. The third generation was marked by engines such as Google which used authority etc. to judge the relevance of the results. The current trend of search engines is to use common databases and directories; hence results from two differnt engines at times may seem very similar.

4 Testing Search Engines

We have been actively involved in the search and discovery problem on the web and in evaluating the quality of the results. The experiments we have conducted are by no means scientific but they do reveal the problems faced by a multitude of people and point out the strengths and weaknesses of search engines. In the following, we give the results of three of these series of tests to compare the results and see the relative effectiveness of the search engines over time. Our conclusions of these tests are given in the concluding section of the paper.

[1] We used these terms since we had a knowledge of all the documents with this terms served by the author's http servers

4.1 First series of tests

Between June 3 and June 15, 1995 the pioneering search engines given in Table 1 were used in a series of tests to find URLs by the author of this paper. Unfortunately, the search engines did not have a method by which to be context sensitive and the searches were made using a target search terms given in section 3. At that time, there were 24 known URLs with the string. These URLs are listed in [12]. All documents in this list existed well before the test date. The results obtained are given in Table 1 giving the number of hits, duplicates, mis-hits and misses. The misses in the results for the manual systems, many of which depended on manual registration of the resources, indicate that the resources have not been registered. As indicated in the results, Lycos was already aggressively crawling the web and hence provided the largest number in the results of the search, most of which were not relevant as indicated in Table1 Figure 1. Yahoo was just starting and they had taken a directory based approach in their search system.

The recall and precision for those search engines producing a non-empty result are given in Figure 1. The highest recall results were about 30%. As shown in Figure 1, the precision of an aggressive crawler such as Lycos was considerably lower at less than 5%.

Search System	Hits	Duplicates	Mis-hits	Missed	Recall %	Precision %
Aliweb	0	0	0	24	0	0
DA-CLOD	0	0	0	24	0	0
EINet	6	0	4	22	8	7
GNA Meta Lib.	0	0	0	24	0	0
Harvest	0	0	0	24	0	0
InfoSeek	7	0	0	17	29	29
Lycos	7	2	222	17	29	3
Nikos	0	0	0	24	0	0
RBSE	0	0	8	24	0	0
W3 Catalog	0	0	0	24	0	0
WebCrawler	4	3	0	20	17	15
WWWW	2	0	0	22	8	8
Yahoo	none	0	0	24	0	0

Table 1. Test results - 1995

4.2 Second Series of tests

Many of the pioneering indexing systems, existing in mid 1995, were no longer accessible when a second series of tests were tried in the fall of 1997. Many of these search engines of the first generation were academic research projects which proved their concepts and the people involved moved on to other challenges. In the meantime, a number of new commercially sponsored systems, such as Altavista, OpenText, Hotbot etc. emerged.

The second series of tests were done in September through October 1997 to find the number of relevant documents that could be located by these search

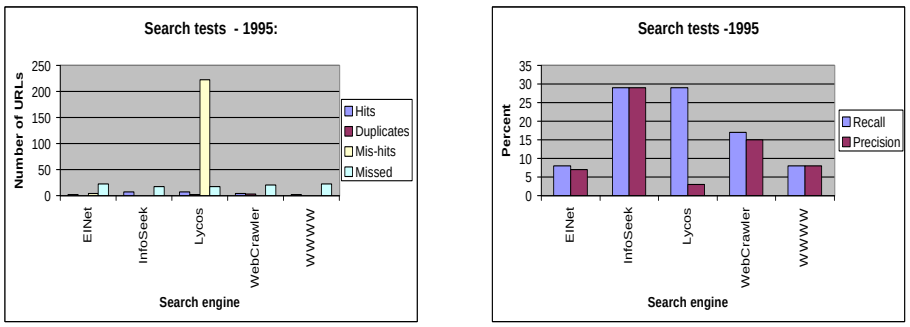

Fig. 1. Tests - 1995: Results(left); Recall and Precision(right)

engines and evaluate the usefulness of the index entries so retrieved[13]. As before, in our tests, the relevance of a document could be judged easily since the set of existing documents was known. We repeated the test performed in 1995 with the same search words. At the time of the tests, some 325 web pages, by the author known to contain the search terms, were accessible on the Web. The complete list of these URLs is given in [14].

Search System	Hits	Duplicates	Mis-hits	Missed	Recall %	Precision %
AltaVista/Yahoo	97	9	23	264	27	25
Excite	114	10	29	247	32	29
Infoseek	8	2	1	319	2	2
Lycos	57	7	15	297	16	15
Hotbot	247	28	58	155	61	51
OpenText	19	-	7	318	6	6

Table 2. Test results - 1997

The test results, given in Table 2, were done on the search engine listed in the table. Note that we used the then recently commercialized version of Lycos in this test. For Web search, Yahoo appeared to use the AltaVista engine and its database and hence produced almost identical results; hence we have given a single result for both search systems in Table 2.

As in the 1995 series of tests, we have given the results by noting the number of hits, the number of duplicates, number of mis-hits and the number of relevant documents missed in the results of the searches. The duplicates are either the same document being served from two sites or same document listed twice. The latter type errors seem to have been corrected in most search engines and they have eliminated such obvious duplicates as we notice from Table 2.

The document missed could be due to the approximations used by engines which use a timeout feature to terminate the search or if parts of the search engine's database is off-line. However, the fact that these search engines could

not locate all relevant documents indicates the inherent problem of the method used in indexing and determining the relevance of the web-pages contents.

The bigger problem is the lack of selectivity and a measure of usefulness of the documents found by the search engines. We have collated the results by following the trail of "next" set of URLs and these could be viewed by pressing on the number of hits for each search engine in the online version[13] of Table 2. A glance at the abstract or summary presented by the search engine was not very informative in judging the relevance of the web page: following the pointers would result in a drain of the searchers time if the page was not very relevant in spite of being placed in the first few pages of the search results.

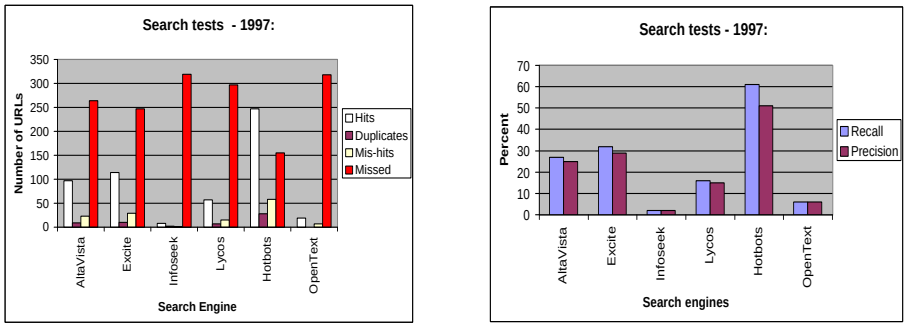

Fig. 2. Tests - 1997: Results(left); Recall and Precision(right)

4.3 Third series of tests

Search System	Hits	Duplicates	Mis-hits	Missed	Recall %	Precision %
AltaVista	99	24	67	230	30	24
Google	155	10	403	174	47	21
HotBot	62	21	121	267	19	13
Lycos	239	37	711	90	73	22

Table 3. Test results - 2001

Table 3 summarizes the test results from a selected number of the third generation of search systems using the same keywords as in the previous tests. The test was carried out in early 2001 [15, 16]. There were 329 web pages containing the search string on the Web well before the tests. The period was judged to be longer than the delay required by most engines before a new web page is indexed. As before, we show the number of hits, duplicates, miss-hits and missed pages for each search engine as well as the recall and precision.

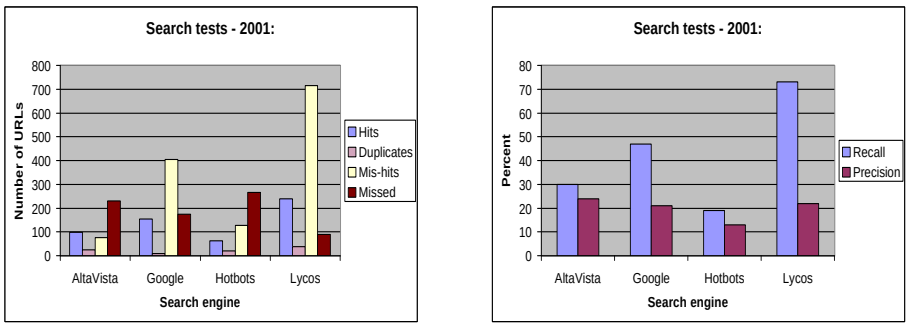

Fig. 3. Test 2001: Results(left); Recall and Precision(right)

Table 3 shows that none of these systems was successful in retrieving all documents sought. The reason for these results is that many of these systems continue to match the specified search terms without regard for the context in which the search words appear in the target information resource. One notes from the results that while the recall percent has increased significantly from the last series of tests, the precision has actually decreased.

Figure 4 gives the distribution of the hits over the default size pages of results from the search engines used in this set of tests. We notice that there are a large number of relevant documents in pages which are considered by the search engines to be of low relevance and hence listed well towards the end. Most users will not have gone past the first few pages; perhaps this is why most engines provide access to the only the top few of pages of the hits (as low as 20 in case of AltaVista). Most typical "surface" users would miss these documents, especially when the user notices a number of pages without any relevant documents as illustrated from the distribution of hits in Figure 4.

5 Conclusion

In Figure 5 we give the percent recall and precision for our search tests over the three generations of search engines. We notice that while the recall has gone up to a very respectable percent for the best, the worst engine still has a recall of less than 20%. Since a large number of hits are in the second half of the search results as depicted in Figures 4, the recall for the casual user, who looks at the first few pages, may not be as high as indicated. We also notice that the precision has actually been reduced from our tests in 1997 and the best value for the tests in 2001 is less than 25%.

Search engines have improved over the years, the number of web pages that these engines have indexed is many hundred millions and in some cases they have exceeded billions. They are extremely useful in tracking the chaotic nature of the web and the pages served; these pages can be in any language and range

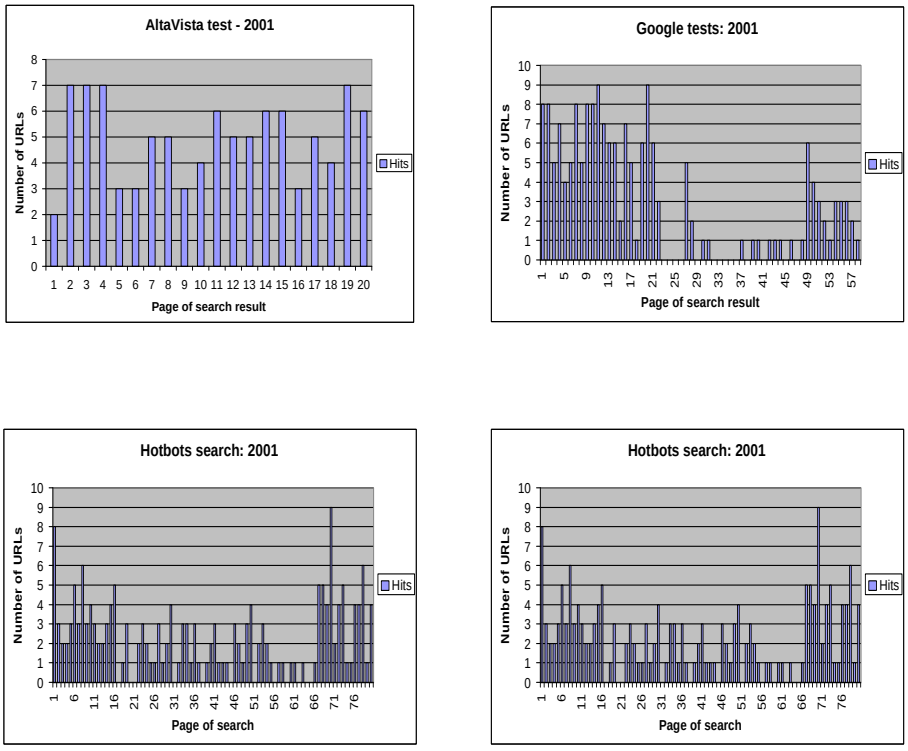

Fig. 4. Test 2001 - hit distribution: AltaVista(top-left), Google(top-right), Hotbot(left); Lycos(right)

from a few lines to megabyte size. It is estimated that over a billion searches are done on these engines daily. They have also improved so that bugs such as "+man +dog -dog -man" which produced non-zero results in earlier systems, now return no result as expected. However, even they can be fooled by a query such as "+chien -dog" with search in any language. Such queries produce results with dogs in it - regardless of whether the search is made on an English or French search engine. Furthermore, some boolean operations still give inconsistent results. Some inconsistencies in the results are temporary others are more chronic.

The problem, raised by the author few years ago[17], with the current automatically generated index databases is that their inadequate semantic information still plagues us. Currently, many search engines allow searches based on terms in fields such as title in the header, URL, etc. and seems like a step in the right direction. However, the relevance used by the search engine may not match that of the user. Judging the relevance of a document is fairly difficult. The relevance based on term frequency, though attractive, has been abused on

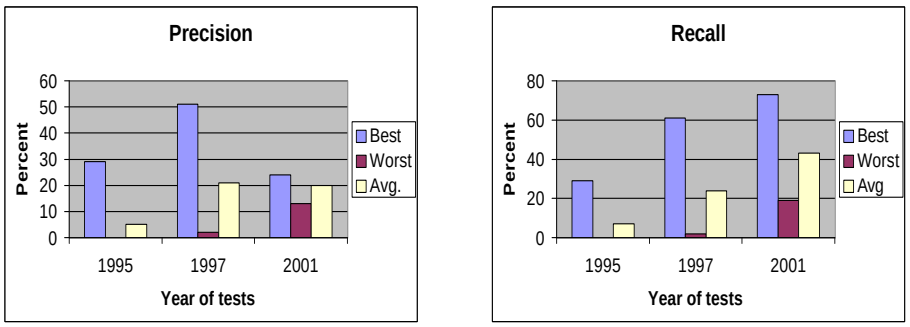

Fig. 5. Tests 1995 - 2001 - trends: Precision(left); Recall(right)

the web and given higher ranking to longer web pages containing the term. Term positioning may be used in ranking but again this has been abused to artificially raise the relevance of a page. With search involving many terms, the proximity of terms may be used to increase the relevance and hence ranking of web pages. Anchor text reference and source authority are now being used by search engines. In anchor text reference, the text that is used as anchor of a web link is used as a term for indexing: if the same text is used by many independent anchors to point to the same web page, the relevance of the web page is increased. This technique is protected from mis-use by assigning higher weights to anchors text from more authorative sources than from lesser ones; thus adding source authority to anchor text. This technique pioneered by Google is now being used by many other search engines. Relevance ranking is constantly being adjusted and improved.

The additional problem faced by users of today's search engine is the paid listing from advertisers which favours the links provided by sponsors of the search engines. Number of hits is not always reported and even if reported not all results are accessible. However, we are completely unaware of what we are missing since most of us would look at the first few pages. With some systems using the first few places to earn revenue, we are duped!

Meta-search engines provide an intermediary system conveniently query many search engines: however this widens the search with the hope that the top hits in one of the target search engines would meet the information needs of the user. Each meta-search engine uses a different set of search engines and method osf processing queries. Some meta-search engines allow user to specify the search engines to use for evaluating the query, while with others the set of engines used is not easy to determine. Some search engines allow the user to specify the time to be used for the search process; others have no such features. The results are again presented by relevance though the relevance used by the meta search engine may not match those of the user. Due to their design, meta search engines are not suitable for an in-depth search since relevant documents which may be

ranked below the threshold of the target search engines would most likely be missed. A search with the "maximum limit simple test" on one of the new meta-search engines for our search string yielded only 10 pages from AltaVista, only a fraction of the one obtained when the search wss made at the site in question.

The other problem with many search engines is that they have stopped giving a count of the number of hits; and when it is given, it is usually incorrect. With the current trend of limiting the search results accessible to the user to only the first few hundred(typically 200), it is unlikely that an ordinary user can repeat our tests in the future.

References

1. Bipin C. Desai, *Report of the Navigation Issues Workshop*, Computer Networks and ISDN Systems, Vol. 27-2, November 1994, pp. 332-333.
2. Robert Caillau, Bipin C. Desai, *Report of the Priorities Workshop*, Computer Networks and ISDN Systems, Vol. 27-2, November 1994, pp. 334-336.
3. Bipin C. Desai, *Report of the Metadat Workshop, Dublin. March 1995.* http://www.cs.concordia.ca/∼faculty/bcdesai/metadata/metadat-workshop-report.html.
4. Bipin C. Desai, Brian Pinkerton, (ed) *Proceedings of the WWW III Workshop on Web-wide Indexing/Semantic Header or Cover Page*, Darmstadt, Germany, April 1995, also available on the Web from
5. Fung R. and Del Favero B. *Applying Bayesian Networks to Information Retrieval, Communication of the ACM, Vol38, No. 3, pp. 42-57, March 1995.*
6. Lewis D., Jones K. *Natural Language processing for information retrieval*, Communications of the ACM, Vol 39, pp. 92-101, January 1996.
7. Sergey Brin, Lawrence Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, http://www-db.stanford.edu/pub/papers/google.pdf
8. Elizabeth Liddy *How a Search Engine Works*, http://www.infotoday.com/searcher/may01/liddy.htm
9. Heather Green *To Google or to GOTO*, http://www.businessweek.com/technology/content/sep2001/tc20010928_9469.htm
10. Gerard Salton, Automatic Information Organization and Retreival, MgGraw-Hill, 1968, New York, NY
11. Katz, W. A. *Introduction to reference Work, Vol. 1-2 McGraw-Hill, New York, NY.*
12. Bipin C. Desai, *Test: Internet Indexing Systems vs List of Known URLs*, http://www.cs.concordia.ca/∼faculty/bcdesai/test-of-index-systems.html
13. Bipin C. Desai, *Test: Internet Indexing Systems vs List of Known URLs: Revisited* http://www.cs.concordia.ca/∼faculty/bcdesai/search-oct97/index.html
14. http://www.cs.concordia.ca/∼faculty/bcdesai/search-oct97/whereis-Desai.html
15. Mohamed Amokrane Mechouet, *Web Based CINDI System*, Masters Thesis, Concordia University, April 2001
16. Bipin C. Desai, *Internet Indexing Systems vs List of Known URLs: 2001 Test Results of Major Search Engines* http://www.cs.concordia.ca/∼faculty/bcdesai/search-2001/index.html
17. Bipin C. Desai, *Supporting Discovery in Virtual Libraries, Jan. 1997*, http://www.cs.concordia.ca/∼faculty/bcdesai