Measuring Search Engine Quality and Query Difficulty: Ranking with Target and Freestyle

Robert M. Losee
School of Information and Library Science
Manning Hall, CB#3360
U. of North Carolina-Chapel Hill
Chapel Hill, NC 27599-3360
losee@ils.unc.edu

and

Lee Anne H. Paris
Oklahoma Christian University Library
Oklahoma Christian University
P.O. Box 11000
Oklahoma City, OK 73136
leeanne.paris@oc.edu

Journal of the
American Society for Information Science
50 (10) 1999, 882–889

July 29, 1999

Abstract

Instead of using traditional performance measures such as precision and recall, information retrieval performance may be measured by considering the probability that the search engine is optimal and the difficulty associated with retrieving documents with a given query or on a given topic. These measures of desirable characteristics are more easily and more directly interpretable than are traditional measures. The performance of the Target and Freestyle search engines is examined and is very good. Each query in the CF database is assigned a difficulty number, and these numbers are found to strongly correlate with other measures of retrieval performance such as an E or F value. The query difficulty correlates weakly with query length.

1 Introduction

Journal articles and conference presentations often describe how one retrieval or search engine was found to be better than another. These comparisons often consist of the presentation of a particular retrieval measure computed for a set of queries that are part of one or more experimental databases. In this study, we introduce a method that can isolate two important factors in the comparison of retrieval results. Below we suggest a method for isolating the intrinsic difficulty associated with using a query and retrieving documents up to a certain point in the search. We also suggest a method for computing the quality of a retrieval or search engine that is independent, in many senses, of the experimental databases used.

The two measures used here differ from those commonly used to evaluate retrieval systems [HH97]. While these measures are computed using some time-consuming techniques that would be virtually impossible to compute with paper and pencil, the results are very easily interpretable: one being a direct measure of query difficulty and the other being a measure of the probability that the search engine will produce optimal ranking. The values of the measures are more easily understood and more directly relevant to the needs of those evaluating retrieval systems than most traditional retrieval performance measures such as precision and recall.

2 Commercial Retrieval Systems and Document Ranking

The Boolean retrieval model has been used by most commercial information retrieval systems since the 1960s, although researchers in the field of information retrieval have suggested a number of other retrieval models, such as the vector space model [SWY75], the probabilistic model [MK60], and the fuzzy set retrieval model [Boo85, SM83]. In 1993, however, two new commercial retrieval engines became available through LEXIS-NEXIS and DIALOG. These retrieval engines, called Freestyle (LEXIS-NEXIS) and Target (DIALOG), have been called natural language search systems because they do not require the user to enter Boolean search statements. As Turtle (1994) stated, a natural language search system "accepts as input a description of an information need in some natural language (such as English) and produces as output a list of documents ranked by the likelihood that they will be judged to match the information need" (p. 212). Even though

Freestyle and Target offer natural language searching, however, neither employs true artificial intelligence, which parses or "understands" a subject domain enough to paraphrase or make conclusions. In fact, the natural language interface in these systems stems from a common automatic indexing strategy that involves three steps:

1. identification of key concepts,
2. removal of stop words, and
3. determination and expansion of root words.

Freestyle and Target are direct competitors and each system has unique aspects. For example, Target eliminates the use of Boolean operators but does not actually process natural language queries. Instead, the system asks users to enter a list of important terms and phrases and then produces a ranked list of documents. A document's rank is based on the number of search terms in the document, the proximity of the search terms to each other in the document, the frequency of a search term in the database, and the length of the document. In order to shed light on how a document's rank is determined, Target provides the frequency of each search term in each document and the relevance weight for each term. Target's creators chose to limit the list of documents retrieved to the 50 documents with the highest ranks, a decision that limits the user's ability to do a comprehensive search in Target.

As far as search aids, Target does not offer a thesaurus, probably because a thesaurus that could be used for all of DIALOG's databases would be have to be painstakingly exhaustive. Target does provide unlimited truncation but not automatic stemming or automatic identification of phrases. Parentheses approximate a Boolean OR, and an asterisk indicates terms that must be present in all documents retrieved. The system defaults to searching for articles published in the past two years, but the desired date of publication can be changed if necessary.

Perhaps the most obvious difference between Target and Freestyle is that Freestyle does allow the user to enter natural language queries, such as "Tell me more about the Gulf War." In fact, Duval and Main (1994) suggest that this feature makes Freestyle particularly appropriate for novice users and users with a vague or ill-defined search topic. The user manual for Freestyle explains that the system is based on "a mathematical concept called associative retrieval" [Mea94, p. 1]. As the manual states:

> Searches using the FREESTYLE feature rely on statistical algorithms that examine your search query, identify and rank relevant search terms and phrases, drop out the "noise" words that won't be searched and compare the relevant terms with every document in the library and file being searched. The FREESTYLE feature then retrieves the top 25 documents that have the best statistical fit with your search terms. [Mea94, p. 1]

The manual goes on to explain that the Freestyle assigns a weight to each query term and then retrieves documents that match the query. A Freestyle query, then, goes through five steps:

1. identification of significant terms and phrases from the query,
2. removal of stop words from the query,
3. calculation of the statistical importance of the terms and phrases in the query and comparison of those terms and phrases to each document in the database,
4. retrieval of documents with the highest probability of matching the query, and
5. ranking of each retrieved document based on the number of query terms in the document and the statistical importance of each query term.

Freestyle also provides date limiting and an online thesaurus. As in Target, mandatory terms may be indicated by an asterisk and a Boolean OR may be indicated through the use of parentheses. The system automatically searches singulars and plurals, but does not offer automatic truncation. Freestyle does recognize more than 300,000 phrases, but the user may indicate unusual phrases (e.g., "fatty acids") by using quotation marks. Relevance feedback is not an option in Freestyle or Target.

The .WHERE and .WHY screens in Freestyle are particularly helpful because they offer some clues about the algorithms used to weight query terms and rank retrieved documents. For example, the .WHERE screen displays a grid that shows the presence or absence of each query term in each document retrieved. On the other hand, the .WHY screen displays the weight assigned to each query term, the number of retrieved documents containing each of the query terms, and the number of documents matched in the database.

LEXIS-NEXIS researchers participated in TREC-2, TREC-3, TREC-4, and TREC-5, but did not employ Freestyle as an information retrieval engine for any of these conferences. Instead, in TREC-3, LEXIS-NEXIS used the SMART system and manual expansion of queries as a retrieval engine, with the result that LEXIS-NEXIS was ranked third in the manually formed ad hoc questions category. In reporting on their TREC-3 system, LEXIS-NEXIS researchers asserted that automatic query expansion is not a "viable option in the on-line service environment because automatic query expansion largely excludes the user from the query formulation process" [LK95]. This argument provides a possible explanation for the lack of relevance feedback in Freestyle. Lu and Keefer also stated that the typical real world query is extremely short, based on the finding that the average length of a Freestyle query was seven terms.

3 Measuring Retrieval Performance

There are a number of ways that retrieval performance may be evaluated. A wide range of performance measuring techniques have been recently summarized by Harter and Hert [HH97] and Burgin [Bur99]. Most of the popular measures assume that documents are either relevant or non-relevant, referred to as binary relevance. There is continuing research on types of relevance [Bar94, Sch94, SW95], how individuals use the concept of relevance [TS98], and performance measures that explicitly allow for continuous relevance [Los98]. We will assume here that documents that are members of a certain set of documents may be referred to as *relevant* and all documents that are not members of this set are referred to as *non-relevant.* We consider the relevance judgments in experimental databases to be approximations of the relevance judgments that might be provided by an actual user.

Performance is most frequently described in terms of precision, the probability that a retrieved document is relevant, and recall, the probability that a relevant document has been retrieved. The performance of a search as it progresses may be shown through use of a precision-recall curve, which shows the qualities of the retrieved set as the search progresses. Precision and recall have been combined into two measures used primarily in the research community, the E and F measures, where $E = 1 - F$, and F is the harmonic mean of the precision and recall measures [VR74, Sha86, SBH97].

Another measure, the *average search length* (ASL), is the average position of relevant documents in the ranked list of documents. A small number repre-

sents superior performance, with the relevant documents moved toward the front of the list of ranked documents. Conversely, a large ASL, more than $N/2,$ where N is the number of documents in the database, represents worse than random performance. When $ASL = N/2,$ performance is random. Related to ASL is Cooper's *expected search length* (ESL), which counts only the non-relevant documents [Coo68]. ESL has an economic interpretation, where non-relevant documents are treated as having a cost when retrieved. The ESL in this case is the average cost associated with retrieving documents, a number that should be minimized by a retrieval system.

In addition to computing the ASL as defined above, we may arbitrarily compute the ASL up to points in the search other than the end of the database. When computing the ASL as above, we may say that the ASL is the average position of a relevant document in the ranked list of the first N documents. Other cutoffs may be used to study retrieval performance up to specific points in the ordered set of documents. A small cutoff might be used to study what is often referred to as a *high-precision* search, while the traditional ASL is essentially the performance using a large part of the full database, a *high-recall* search.

4 Analytic Models of Performance

The performance of a retrieval or filtering system may be computed analytically using probabilistic methods, with probabilistic parameters as input, rather than using repeated experimentation from which performance results are extrapolated [Los98]. Given the set of characteristics of a set of documents, the performance may be directly computed. The analysis here computes the parameters associated with ranked lists of documents from searches producing ranked output using commercial search engines. Using this model can lead to an understanding of the direct relationship between retrieval performance and changes in queries, documents, relevance judgments, database size, and document rankings.

Following the development in Losee [Los98], the average search length, for the case of a single binary feature, is

$$ASL = N\left[QA + (1-Q)(1-A)\right] \tag{1}$$

where N is the number of documents in the database, Q is the probability that ranking is optimal (the *quality* measure), and A is the expected proportion of documents examined in an optimal ranking if one examines all the documents up

to the document in the average position of a relevant document, or the optimal ASL re-scaled to the range from 0 to 1. We compute

$$A = \frac{1 - p + t}{2}, \tag{2}$$

where p is the probability that the feature occurs in relevant documents and t is the unconditional probability that the feature occurs.

When $A = 0,$ for example, the relevant documents would be at the front of the ordered list of documents, while when $A = .5,$ the average position for a relevant document is in the middle of the ordered list of documents. Interestingly, a very bad A value, such as $A = 1,$ works well with a worst-case retrieval engine, with $Q \rightarrow 0.$ This is equivalent to placing the best documents at the end of the list of ranked documents ($A = 1$) and then retrieving them backwards ($Q = 0.$) When discussing results below, only the positive results are given, with those cases where a "double negative" produces a positive result being ignored, and the twin positive-positive process, producing a positive retrieval result, being presented instead.

We will use this simple retrieval model to capture the performance of the search engines and queries being studied. If we consider the queries as each representing an information need or concept cluster, we may treat this concept cluster as a single feature upon which we desire to discriminate. This cluster also has probabilities of occurring in various classes of documents. One could examine the number of terms in queries and attempt to model the system using this many Q values, which might then be averaged in some way, but accurate estimates would require far greater numbers of documents with relevance judgments than are typically available in experimental databases. We believe that using this single concept model is an adequate approximation of what would be obtained with a multivariate model, where accurate estimates would require very large numbers of document rankings.

5 Document Rankings from Commercial Retrieval Engines

Any experiments involving Freestyle and Target must of necessity be "black box" experiments [RHB92], since the algorithms used in these retrieval systems are trade secrets. Based on system documentation, however, we can conclude that

both systems employ algorithms based on the vector space and probabilistic models, although the exact values used to calculate relevance remain a mystery. In their evaluation of Target, Tenopir and Cahn [TC94] state that document weights are adjusted for document length, but Keen [Kee94] asserts that he did not detect any clear evidence of such adjustment. Ingwersen [Ing96] suggests that "Target is applying quorum logic (in the traditional way), document term frequencies and collection term frequencies as elements of its ranking algorithm," (p. 45) but provides no evidence for this claim. We do know, however, that Target's ranking algorithm includes at least four variables [Kee94]:

1. number of search terms in each record,
2. proximity of search terms to each other in a record,
3. frequency of a term in the database, and
4. length of the document.

Freestyle, on the other hand, provides a little more information about the information retrieval process used. For example, the .WHERE and .WHY screens in Freestyle show that a term's weight is inversely proportional to its frequency in the database. In fact, the Freestyle HELP explanation about query term weights states that "term importance is based on how frequently the term appears in the file(s) you are searching. The more often a term occurs, the lower its term importance." These facts, then, suggest that the system employs some version of inverse document frequency to calculate term weights [SJ72]. When calculating the inverse document frequency weight (IDF), "terms with medium to low collection frequencies are assigned high weights as good discriminators, while frequent terms have low weights" [RSJ76, pp. 129-30]. The matching algorithm for Freestyle appears to be derived from the vector space and probabilistic models, where the weight of each document is the sum of the products of term weights and frequencies of the terms in the document. The ranking algorithm for Freestyle, then, appears to involve a minimum of three variables:

1. frequency of a search term within the database,
2. frequency of a search term in a record, and
3. number of search terms in a record.

5.1 Experimental Rankings

A series of document rankings were obtained and then analyzed to determine both the retrieval performance of different search mechanisms and the difficulty or quality of individual queries and topics. The document rankings were from the subsets of documents on the appropriate system that contained the medical subject heading CYSTIC FIBROSIS (CF) in the MEDLINE database during the period from 1974 to 1979 [SWWT91, Par98, PT98]. The original CF queries were produced by subject specialists and are described in Shaw et al. [SWWT91]. Different forms for the queries have been produced by Tibbo, as described in Paris and Tibbo [PT98]. Exhaustive relevance judgments have been obtained for this data, making it an attractive subset of MEDLINE for studies of retrieval performance.

Paris [Par98] developed the six sets of document rankings referred to below as *Boolean, Freestyle1, Freestyle2, Freestyle3, Target1* and *Target2*. The first, referred to below as *Boolean*, represents the retrieval performance obtained with a set of Boolean queries developed by Helen Tibbo for her work with the CF database. In her study, Tibbo found which of several forms for a query produced the best results, and this form of the query is used in this study.

Freestyle1 and Target1 represent document rankings produced by searches on the corresponding system using terms from the original natural language queries. Freestyle2 and Target2 represent queries constructed from the terms used in the optimal queries developed by Helen Tibbo. For Freestyle2 and Target2, terms were placed in a single set of quotes if they were linked using the adjacency operator in the optimal Boolean query, e.g. *information ADJ retrieval* would become "information retrieval." Terms connected by an OR were placed in parentheses, which are used to approximate OR in Target and Freestyle. Freestyle3 contains the full natural language form of the query, e.g., "What are the hepatic complications or manifestations of CF?"

Minor changes that had to be made in a few specific cases due to system limitations are described in Paris [Par98]. Of greatest interest here is that Freestyle rankings are limited to 1000 documents, while those of the Boolean system includes all 1239 documents in the CF database. Target retrieves up to 50 documents.

Note that our examination of these commercial search engines, as well as some earlier studies, are based on retrieval using titles, abstracts, and the sophisticated controlled vocabulary used by the National Library of Medicine. Those search engines using techniques optimized for full-text retrieval will perform somewhat differently with entire documents than they do with the CF database.

6 System Quality and Query or Subject Difficulty

Given the analytic model of retrieval, we may compute the A values for each query (A_i represents the A value for the i^{th} query) and the Q for each retrieval engine, where Q_j represents the quality (probability of optimal ranking) of search engine j. The A values may be interpreted as the level of difficulty associated with retrieving the relevant documents on the topic represented by various formulations of the query. The Q values may be interpreted as the quality of each search mechanism.

We compute these values by performing a rather lengthy regression. Our goal is to solve for the various values of A_i and Q_j for each query and each search engine, finding the set of A and Q values that minimize the errors made in estimating the ASL values. This is a complex problem, and there are no standard simple procedures for solving it. We can treat the problem as being to solve a non-linear regression of the form

$$\begin{aligned} ASL = N\Big[&(x_1A_1 + x_2A_2 + \cdots + x_{100}A_{100}) \\ &(y_1Q_1 + y_2Q_2 + \cdots + y_6Q_6) \\ &+ (1 - (x_1A_1 + x_2A_2 + \cdots + x_{100}A_{100})) \\ &(1 - (y_1Q_1 + y_2Q_2 + \cdots + y_6Q_6))\Big]. \end{aligned}$$

Here the ASL is the dependent variable and the parameters $Q_1, Q_2, \ldots, Q_6$ and $A_1, A_2, \ldots, A_{100}$ are independent variables to be estimated by the regression package. The variable x_i is an indicator variable that has the value 1 when the query in question is query i, and 0 otherwise. The variable y_i similarly is an indicator variable that has the value 1 when the retrieval engine being used is retrieval engine number i, and 0 otherwise. The data set contains 600 document rankings, one for each combination of the six search techniques and for each of the 100 queries. The N values are set to the correct number of documents for each database.

The numbers that are obtained from these regressions are inexact. They are estimates that would be better with a larger sample of queries and documents from which to make the estimates. The standard errors for estimating Q values are all approximately 0.014, while the standard errors for estimating A values are approximately 0.056.

The Q values reflect the database from which they are derived. The A values are query specific and reflect the nature of the relevance judgments and the documents available. The Q values are computed so as to mathematically complement

		Q_i *for queries*		
i	*Engine*	$1-100$	$1-50$	$51-100$
1	Boolean	1.000	1.000	1.000
2	Freestyle1	0.871	0.868	0.873
3	Freestyle2	1.000	1.000	1.000
4	Freestyle3	0.913	0.919	0.907
5	Target1	0.744	0.751	0.738
6	Target2	0.956	0.927	0.983

Table 1: Q values for full retrieval for the 6 different retrieval engines.

the A values so the regression formula produces an ASL values with minimal error. While Q values clearly will vary due to the characteristics of a specific database, the variance should be relatively small compared to the variation obtained with other measures of retrieval performance quality, such as precision. In the following section we examine the Q values and their robustness.

7 Comparing Retrieval or Search Engines

The data in Table 1 contain the set of Q values that are obtained when full retrieval is used, with no cutoffs. We notice that the values for Boolean and for Freestyle2 show that the engines appear to be optimal at this point, with Target2 being a somewhat lower performing engine. The columns on the right side of Table 1 show the Q values for the first and the second 50 queries, showing the relative robustness of the Q values.

The quality or difficulty associated with retrieving documents for each query is provided by the query-specific scores provided in Table 2. Interestingly, some queries show that relevant documents are easily moved to the front of the list of ranked documents (e.g. queries 9 through 11) while the A values for other queries, such as 8 and 24, represent situations where it is far more difficult to discriminate between relevant and non-relevant documents.

These numbers may be interpreted easily by noting that there are about 1000 documents being considered for retrieval. If each A value is multiplied by 1000, we obtain the expected position of a relevant document if ranking is optimal. For query 1, $A_1 = .02$, suggesting that the average position of a relevant document

Query i	A_i	*Query* i	A_i	*Query* i	A_i	*Query* i	A_i
1	.020	26	.075	51	.345	76	.000
2	.500	27	.048	52	.000	77	.087
3	.087	28	.000	53	.000	78	.199
4	.064	29	.240	54	.044	79	.080
5	.102	30	.000	55	.187	80	.002
6	.027	31	.074	56	.000	81	.104
7	.122	32	.023	57	.076	82	.106
8	.326	33	.109	58	.151	83	.165
9	.000	34	.142	59	.279	84	.000
10	.000	35	.106	60	.094	85	.123
11	.000	36	.022	61	.047	86	.153
12	.054	37	.141	62	.226	87	.101
13	.196	38	.200	63	.169	88	.000
14	.092	39	.118	64	.045	89	.061
15	.147	40	.248	65	.155	90	.000
16	.220	41	.012	66	.107	91	.346
17	.034	42	.169	67	.162	92	.068
18	.060	43	.123	68	.324	93	.056
19	.101	44	.150	69	.000	94	.194
20	.043	45	.083	70	.000	95	.075
21	.000	46	.035	71	.000	96	.061
22	.149	47	.106	72	.032	97	.000
23	.228	48	.053	73	.000	98	.089
24	.318	49	.026	74	.055	99	.000
25	.132	50	.069	75	.000	100	.500

Table 2: The set of A values obtained with retrieval of all documents.

would be at about the 20th document retrieved. A query such as 24, where $A_{24} = .318$, suggests that the average position of a relevant document would be at about the 318^{th} document retrieved. Clearly, most searchers who want high-recall will find an A value for this dataset over $.1$ to be unacceptable.

The easy interpretation of measures such as A is one of the reasons for its use. If A remains constant, e.g. $A = .001$, one can easily see the practical impact for the searcher of retrieving these documents from a database of a thousand, a million, or hundreds of millions of documents.

8 Performance Superiority over a Range of Situations

Retrieval performance may be measured at a single point, as was done above, or performance might be measured over a range of values for a variable. For example, examining the performance characteristics over a range of cutoffs from 2 to N, the number of documents in the database, can show how a given search engine performs at different points in the search process. A search engine optimized for high-precision lower-recall searches, for example, might have a higher Q at cutoff 10 than at cutoff N. These variations may be computed experimentally, computing performance at individual cutoff points, or analytically, showing through proof methods that one retrieval engine is superior to another over a range of cutoffs or other values.

Figure 1 shows the Q values computed for 4 retrieval engines. There is clearly noise in the figures for low cutoffs, with trends only appearing with higher cutoffs, where larger amounts of data are available for computation.

Figure 2 shows in more detail the Q values for low cutoffs. The Target 2 retrieval engine, which performed very well for larger cutoffs, shows a lower level of performance for small cutoffs, suggesting that it is probably better for higher recall searches.

9 Query Difficulty and Correlates with Other Performance Characteristics

The difficulty associated with an individual query, A, may be compared to other query-specific performance figures in an effort to validate the use of the A mea-

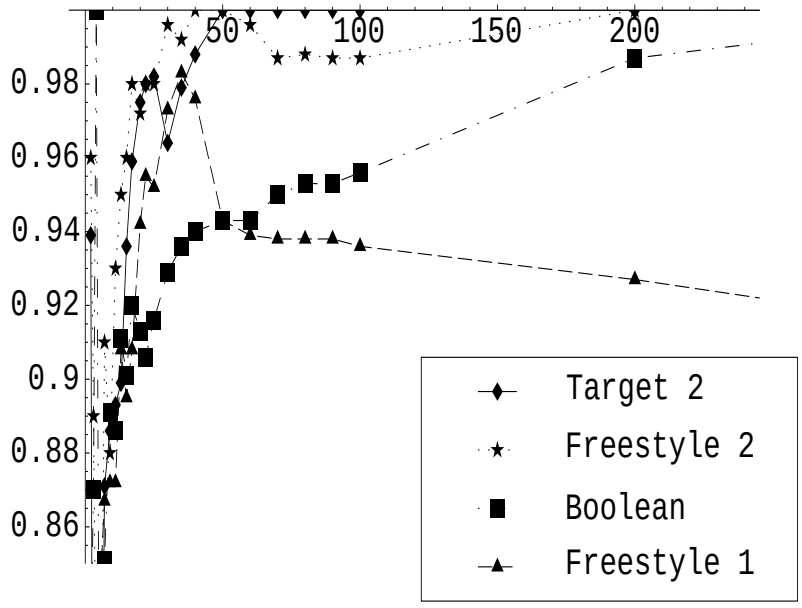

Figure 1: Q values at varying cutoffs for different retrieval engines.

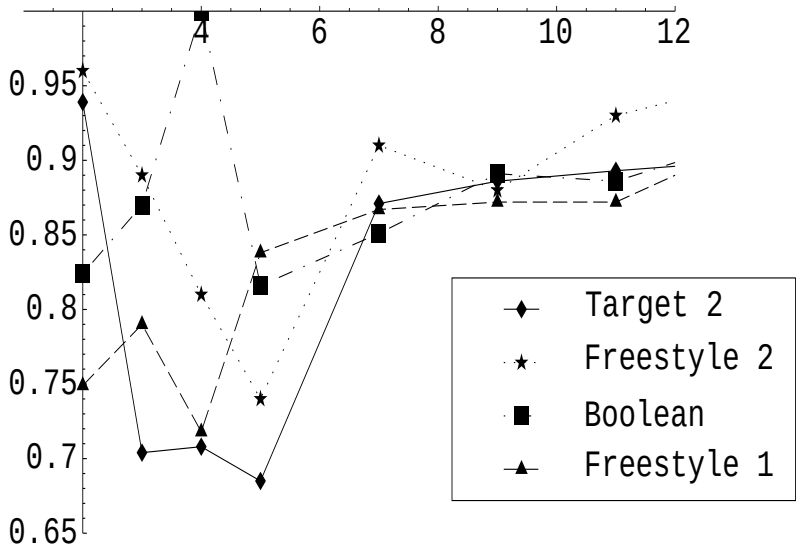

Figure 2: Q values at varying cutoffs for different retrieval engines.

sure. A strong correlation between the measures would support the validity of the proposed A measure. While a correlation between A and a measure M may show a relationship, it does not necessarily imply that A is measuring the same phenomenon as M.

In Paris and Tibbo [PT98], a set of *E* values are reported that were correlated in our study with A values. The E values were obtained at the highest recall available for that particular query from the CF database. An E value was unreported for query number 2 which had no relevant documents. We conservatively chose to use the worst-case E value (1) as the E value for this query in this study. The Spearman rank correlation between the A values and the E values is $.523$, and the Pearson product moment correlation is $.407$. We may interpret these strong correlations as indicating the degree to which the value of a traditional performance measure such as E is due to the difficulty of the individual queries.

There is a positive correlation between the A values and the number of natural language terms in the query, with the Pearson correlation being $.172$ and the Spearman rank correlation being $.126$. This suggests that shorter queries produce better results than do longer queries, which is contrary to the idea that the increased richness obtained with longer queries makes up for the additional noise created by adding terms. Several factors may be at work here. Some of the longer queries include details about what the searcher wants, for example, the clause at the end of query 34, "... what are their relative advantages and disadvantages?" Query 37 adds a second question "... and what factors contribute to erroneous results of these tests?" These longer queries express information needs that are inherently more abstract and are less topical. They add little to the performance of a term-matching or weighting search engine, although these additional clauses are certainly helpful to human searchers in developing queries and evaluating documents.

The correlation between the number of terms from a public domain medical dictionary and the A values was negligible, suggesting that query difficulty isn't simply a matter of adding or deleting sublanguage terms from natural language queries. The unnamed machine readable medical dictionary was obtained from the PC-SIG library of public domain software (Disk 4160, 13th edition, CDROM version) and was manually supplemented to include most of the specialized medical terms found in the CF database.

10 Discussion and Recommendations

The work here has addressed the question of the relative performance of several different retrieval engines and systems, as well as the difficulty associated with retrieving documents for specific queries or topics. We have developed a technique for estimating the probability of optimal ranking for a retrieval engine, allowing us to isolate this value which characterizes the quality of a search engine from the query-retrieval difficulty, associated with retrieving the specific query and the documents relevant to the query. These query-specific A values correlate with other performance measures, such as E, providing empirical support for the usefulness of A.

The results suggest that *slightly* better subject-based retrieval performance is obtained with best-case Boolean searching or the ranking engine used by Freestyle when compared to the ranking engine used by Target.

Sembok and Van Rijsbergen [SV90] noted, before the introduction of Target or Freestyle, that "the keyword approach with statistical techniques has reached its theoretical limit and further attempts for improvement are considered a waste of time." While this statement may be a bit strong, there is little difference between the two commercial search engines in terms of performance, despite commercial pressures to develop a better search engine, and this performance may be about the best obtainable without using much more sophisticated techniques and knowledge, that is, without revolutionary changes in retrieval theory or practice.

The research discussed here has been based on tests using the CF dataset, described in Shaw et al. [SWWT91]. This dataset has exhaustive relevance judgments and is thus an excellent database for many research purposes. While the CF database can be used in experimental systems, the same set of documents also can be retrieved from existing commercial systems, making the dataset invaluable for the study of commercial system performance. However, full-text systems containing entire documents, instead of just titles, abstracts, and descriptors, can be expected to perform somewhat differently, and this study provides only an approximation of the performance that would be obtained with retrieving full documents using these particular commercial search engines.

Future research might address further aspects of the query and how its characteristics affect performance. By computing the correlations between A and other factors, we can look at measures of query-specific retrieval difficulty and other factors that may cause the query to be effective or ineffective at separating the documents that the user considers to be relevant from those documents the user considers to be non-relevant. We may also consider more elaborate analytic mul-

tivariate approaches to the study of retrieval performance. Given much larger sets of documents, multivariate techniques [Los98] can be used to more accurately estimate both the performance of different search engines and the query difficulty due to specific characteristics of a query.

References

[Bar94] Carol L. Barry. User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45(3):149–159, 1994.

[Boo85] Abraham Bookstein. Probability and fuzzy-set applications to information retrieval. In Martha Williams, editor, *Annual Review of Information Science and Technology*, volume 20, pages 117–151. American Society for Information Science, Washington, D.C., 1985.

[Bur99] Robert Burgin. The Monte Carlo method and the evaluation of retrieval system performance. *Journal of the American Society for Information Science*, 50(2):181–191, 1999.

[Coo68] William S. Cooper. Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science*, 19(1):30–41, 1968.

[HH97] Stephen P. Harter and Carol A. Hert. Evaluation of information retrieval systems: Approaches, issues and methods. In Martha Williams, editor, *Annual Review of Information Science and Technology*, volume 32, pages 1–94. American Society for Information Science, Washington, D.C., 1997.

[Ing96] Peter Ingwersen. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1):3–50, 1996.

[Kee94] E. M. Keen. How does Dialog's Target work? *Online and CDROM Review*, 18(5):285–288, 1994.

[LK95] X. A. Lu and R. B. Keefer. Query expansion/reduction and its impact on retrieval effectiveness. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 231–239. National Institute of Standard and Technology, Computer Systems Laboratory, Gaithersburg, MD, 1995.

[Los98] Robert M. Losee. *Text Retrieval and Filtering: Analytic Models of Performance*. Kluwer, Boston, 1998.

[Mea94] Mead Data Central. *A Handbook for Using Lexis/Nexis*. Mead Data Central, New York, 1994.

[MK60] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing, and information retrieval. *Journal of the ACM*, 7:216–244, 1960.

[Par98] Lee Anne H. Paris. *Stopping Behavior: User Persistence in Online Searching and Its Relation to Optimal Cutoff Points for Information Retrieval Systems.* PhD thesis, University of North Carolina, School of Information and Library Science, Chapel Hill, 1998.

[PT98] Lee Anne H. Paris and Helen R. Tibbo. Freestyle vs. Boolean: A comparison of partial and exact match retrieval systems. *Information Processing and Management*, 34(2/3):175–190, 1998.

[RHB92] Stephen E. Robertson and M. M. Hancock-Beaulieu. On the evaluation of IR systems. *Information Processing and Management*, 28(4):457–466, 1992.

[RSJ76] Stephen E. Robertson and Karen Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976.

[SBH97] William M. Shaw, Jr., Robert Burgin, and Patrick Howell. Performance standards and evaluations in IR test collections: Vector-space and other retrieval models. *Information Processing and Management*, 33(1):15–36, 1997.

[Sch94] Linda Schamber. Relevance and information behavior. In Martha E. Williams, editor, *Annual Review of Information Science and Technology*, volume 29, pages 3–48. American Society for Information Science, Washington, D.C., 1994.

[Sha86] William M. Shaw, Jr. On the foundation of evaluation. *Journal of the American Society for Information Science*, 37(5):346–348, 1986.

[SJ72] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[SM83] Gerard Salton and Michael McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[SV90] T. Sembok and C. Van Rijsbergen. SILOL: A simple logical-linguistic retrieval system. *Information Processing and Management*, 26(1):111–134, 1990.

[SW95] Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition*. Blackwell, Oxford, second edition, 1995.

[SWWT91] William M. Shaw, Jr., Judith B. Wood, Robert E. Wood, and Helen R. Tibbo. The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research*, 13:347–366, 1991.

[SWY75] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[TC94] C. Tenopir and P. Cahn. Target and Freestyle: Dialog and Mead join the relevance ranks. *Online*, 18(3):31–47, 1994.

[TS98] Rong Tang and Paul Solomon. Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. *Information Processing and Management*, 34(2/3):237–256, 1998.

[VR74] C.J. Van Rijsbergen. Foundation of evaluation. *Journal of Documentation*, 30(4):365–373, 1974.