Research Paper ■

Using Citation Data to Improve Retrieval from MEDLINE

ELMER V. BERNSTAM, MD, MSE, JORGE R. HERSKOVIC, MD, MS, YINDALON APHINYANAPHONGS, MS, CONSTANTIN F. ALIFERIS, MD, PHD, MADURAI G. SRIRAM, PHD, WILLIAM R. HERSH, MD

Abstract **Objective:** To determine whether algorithms developed for the World Wide Web can be applied to the biomedical literature in order to identify articles that are important as well as relevant.

Design and Measurements: A direct comparison of eight algorithms: simple PubMed queries, clinical queries (sensitive and specific versions), vector cosine comparison, citation count, journal impact factor, PageRank, and machine learning based on polynomial support vector machines. The objective was to prioritize important articles, defined as being included in a pre-existing bibliography of important literature in surgical oncology.

Results: Citation-based algorithms were more effective than noncitation-based algorithms at identifying important articles. The most effective strategies were simple citation count and PageRank, which on average identified over six important articles in the first 100 results compared to 0.85 for the best noncitation-based algorithm ($p < 0.001$). The authors saw similar differences between citation-based and noncitation-based algorithms at 10, 20, 50, 200, 500, and 1,000 results ($p < 0.001$). Citation lag affects performance of PageRank more than simple citation count. However, in spite of citation lag, citation-based algorithms remain more effective than noncitation-based algorithms.

Conclusion: Algorithms that have proved successful on the World Wide Web can be applied to biomedical information retrieval. Citation-based algorithms can help identify important articles within large sets of relevant results. Further studies are needed to determine whether citation-based algorithms can effectively meet actual user information needs.

■ **J Am Med Inform Assoc.** 2006;13:96–105. DOI 10.1197/jamia.M1909.

Information overload is no longer a theoretical issue in biomedicine but a real impediment to education, research, and clinical care. According to National Library of Medicine (NLM) figures, a physician reading two articles daily would be 550 years behind within one year.[1] As the literature grows, so does the number of articles containing a given search phrase, especially for broad topics. For example, a PubMed search for [tamoxifen AND ''breast cancer''] retrieves 6,750 articles, too many for a human to easily review.[2] Therefore, we must develop information retrieval strategies to identify articles that are important as well as relevant.

Like the biomedical literature, the World Wide Web (WWW) is large and growing rapidly. Since the mid-1990s, researchers in academia and industry have recognized the difficulty and importance of information retrieval from the WWW. Consequently, a tremendous amount of research has gone into developing strategies for effective retrieval of information from hyperlinked environments. Our unifying hypothesis is that successful techniques developed for the WWW can be adapted to help users access the biomedical literature more effectively. On the WWW, search algorithms that make use of link information have proven successful. We draw an analogy between links from one Web page to another and citations (references) from one article to another. Therefore, we hypothesize that citation-based algorithms will be more effective than noncitation-based algorithms at identifying important articles.

Affiliations of the authors: School of Health Information Sciences, The University of Texas Health Science Center at Houston, Houston, TX (EVB, JRH, MGS); Department of Biomedical Informatics, Vanderbilt University, Nashville, TN (YA, CFA); Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR (WRH).

Supported in part by NLM grant 5 K22 LM008306 and a training fellowship from the W. M. Keck Foundation to the Gulf Coast Consortia through the Keck Center for Computational and Structural Biology.

The authors are also grateful to Thomson-ISI for granting use of the Science Citation Index for research purposes.

Correspondence and reprints: Elmer Bernstam, MD, School of Health Information Sciences, The University of Texas Health Science Center at Houston, 7000 Fannin Street, Suite 600, Houston, TX 77030; e-mail: <elmer.v.bernstam@uth.tmc.edu>.

Received for review: 07/12/05; accepted for publication: 09/16/05.

Background

Searching the Biomedical Literature

MEDLINE, created and maintained by the NLM, is the world's premier bibliographic database of biomedical literature and is available via multiple interfaces including PubMed, which is also maintained by the NLM. Between 1999 and 2002, the number of PubMed searches increased 50% from 244 million/year to 380 million/year. Over the same time period, the number of articles indexed per year grew 13.6% to over 502,000 in 2002 alone.[3]

MEDLINE searching can affect care and perhaps even improve clinical outcomes.[4–7] Multiple studies have assessed MEDLINE information retrieval. Many found ineffective retrieval, especially for queries issued by novice users. A comprehensive review is beyond the scope of this article, and the interested reader is referred to the latest systematic review.[8] Our work is motivated by a desire to improve care and to enhance research by more effective access to the biomedical literature.

Recognizing the need to keep up with the literature, starting in 1992 the NLM began producing bibliographies of relevant articles on a variety of topics.[9] Bibliographies contain thousands of references that are compiled manually. For example, at the time of this writing, the most recent bibliography addressed celiac disease and contained 2,382 references covering the period 1986 to 2004. Systematic reviews such as the NLM bibliographies require a great deal of human effort. Human effort is required initially to compile the bibliography and then to keep the bibliography up-to-date. If important articles could be identified automatically or semiautomatically, systematic reviews would be easier to compile and maintain.

Relevance, Quality, and Importance

Traditional information retrieval evaluations rely on the concept of relevance, which, although difficult to define precisely, refers to the question of whether a search result deals with the same concepts as the query. Alternatively, a relevant article satisfies the information need of the user who issued the query. An information need is the user's expression, in his or her own language, of the information that he or she desires.[10] A good information retrieval strategy returns all relevant documents (high recall*) and only the relevant documents (high precision†). However, as literature databases grow, a high recall/high precision strategy may still produce a very large result set.

To help focus users' attention on articles that are most likely to be useful, researchers developed quality filters that return relevant articles that also conform to methodological quality standards. However, even quality filters tuned for precision rather than recall retrieve thousands of articles about common conditions. Clinical query templates based on the work of Haynes et al. are described below and are available on the PubMed Web site.[11] The specific, high precision, query template for therapy returns over 3,800 results for "breast cancer" (query performed on June 14, 2005), far too many for a human to review. The sensitive, high recall version of the same query template returns over 40,000 results.

Importance refers to an article's influence (or predicted influence) on the scientific discipline. In order to identify the "must-read" articles on a particular topic, a user could first retrieve relevant articles, then filter using measures of quality, and finally prioritize them in decreasing order of importance. One way to operationalize importance is to use citation analysis. A highly cited article has affected the field more than an article that has never been cited. Therefore, it may be reasonable to test citation-based algorithms with respect to their ability to identify important articles. However, we note that some authors question the relationship between citation analysis and importance.[12] This framework is partly based on an analogous discussion of information retrieval on the WWW.[13]

Methodological quality and importance are related, but not interchangeable. A high-quality article is not necessarily important nor is an important article necessarily of high quality. For example, some areas of biomedicine are not amenable to randomized controlled trials. Such areas may include disciplines where sham operations or even randomization may not be ethically permissible. Consequently, a case series may be very important in surgical oncology, but not in hypertension. Further, biological literature cannot be judged by the standards of evidence-based medicine (EBM) that form the basis of clinical query templates and machine learning models discussed in this paper. Therefore, metrics that do not rely on methodological quality may be complementary to existing methods.

Retrieving High-Quality Articles (Clinical Query Templates)

MEDLINE query templates are relatively complex generalized queries that were systematically constructed by humans and validated against a manual review. Query templates are an attempt to "bottle" search expertise for the benefit of novice users. Query templates take advantage of MEDLINE features, such as publication types, that are not familiar to novice users. For example, retrieving randomized controlled trials rather than opinion pieces. Although query templates can effectively retrieve high-quality articles, results are generally not ordered by importance or quality. For example, PubMed clinical query templates retrieve results in reverse chronological order.[14,15] In spite of this limitation, multiple MEDLINE interfaces have implemented query templates.[15] PubMed clinical queries using research methodology filters are available on the PubMed Web site.[16]

Retrieving High-Quality Articles (Machine Learning)

Machine learning algorithms such as support vector machines (SVMs) attempt to automatically identify features that distinguish desirable articles. Given adequate training data consisting of positive and negative examples, machine learning requires less human effort to develop effective search strategies. Machine learning approaches based on SVMs perform well and can rank articles in order of quality. A study by Aphinyanaphongs et al.[17] found that SVM models outperform clinical query templates when using the *American College of Physicians Journal Club* (*ACP-JC*) as a gold standard. The authors used a robust cross-validation method to estimate generalization error but did not apply their models to other document sets, domains, and/or gold standards.

Using Citations to Rank Search Results

Currently available MEDLINE interfaces rank results by similarity to the query (MDConsult®, St. Louis, MO), publication date (PubMed, MDConsult®, and Ovid®, New York, NY), availability of full text (MDConsult®), or MEDLINE record information such as ID number. These ranking methods do not make use of human judgments inherent in citations. However, experience on the WWW suggests that citations may be a powerful way of improving a search. Google™'s PageRank[18] and Kleinberg's HITS[19] (Hyperlink-Induced Topic Search) algorithms pioneered the use of Web links, or references from one page to another. PageRank has proven so successful that most Web search engines now use link-based algorithms.[20]

Implicit in this paper is the theme of endowing automated systems with knowledge about the biomedical literature. For example, combining the Science Citation Index (SCI®), a database of citations, with MEDLINE is routinely and effectively used by medical librarians. Novice users, however, may not know how to search using multiple databases or

*Recall = percentage of relevant articles contained in the database that are retrieved.

†Precision = percentage of retrieved articles that are relevant.

Table 1 ■ Algorithms

Algorithm	Category	Description
Clinical queries (sensitive)	Retrieval	PubMed clinical queries based on Haynes et al.[11]
Clinical queries (specific)	Retrieval	
Simple PubMed queries	Retrieval	Simple queries intended to simulate a naïve PubMed user (please see Appendix 1, available as a *JAMIA* online supplement at http://www.jamia.org, Tables A2 and A3).
Zettair	Retrieval	Public domain information retrieval system based on a vector space model (TF*IDF) and is available from the Search Engine Group at the Royal Melbourne Institute of Technology.[36]
ML-EBM	Retrieval	Articles ranked by machine learning models built to identify articles matching an EBM standard.[25]
Impact factor	Ranking	Articles ranked by impact factor of the journal in which the article was published (2003 edition available from Thomson-ISI, Stamford, CT).[23]
Citation count (CC)	Ranking	Articles ranked by the number of citations to the article based on the Science Citation Index (updated through 2004 Q4, Thomson-ISI, Stamford, CT).[37] Articles with the largest number of citations are listed first.
PageRank	Ranking	Articles ranked using PageRank.[38]

ML-EBM = Machine Learning—Evidence-Based Medicine.

when this is beneficial. Further, citation-based ranking of large result sets using data derived from the SCI is not practical using currently available tools.

Methods

Gold Standard (Society of Surgical Oncology Annotated Bibliography)

The Society of Surgical Oncology (SSO), a professional organization, maintains an "annotated bibliography of important literature on common problems in surgical oncology" (SSO-AB). The SSO-AB is a collection of 457 unique MEDLINE-cited articles, with additional content from meeting abstracts and clinical trial descriptions. In this study, we consider only the MEDLINE-cited articles. Articles are grouped into ten categories: breast cancer, colorectal cancer, endocrine oncology, esophageal cancer, gastric adenocarcinoma, hepatobiliary malignancies, lung cancer, melanoma, pancreas cancer, and soft tissue sarcoma. Each category was compiled by a single expert (editor) and reviewed by a panel of experts on that particular topic (i.e., ten panels).

The bibliography is available from the SSO Web site at http://www.surgonc.org. Although the SSO intended to update the SSO-AB yearly, the latest (second) edition is dated October 2001. The first edition, published in 2000, contained 381 MEDLINE-cited articles. Of the 457 articles in the second edition, 310 were also present in the first edition. As of this writing, it has not been updated in over three years, reflecting the amount of human effort required to manually maintain an annotated bibliography. For the purposes of our study, we considered articles that are referenced by the SSO-AB to be "important." Articles that are not included in the SSO-AB may also be important, but the SSO-AB articles were chosen by experts as the most important or "must read." Therefore, SSO-AB articles should be returned at the top of result sets.

Information retrieval research has generally relied on human expert opinion regarding the relevance of retrieved articles. For example, OHSUMED, a widely used MEDLINE test collection, contains articles associated with human relevance judgments.[21] Ideally, panels of experts rather than individuals evaluate search results. However, expert time is precious and panels are often impractical. Since the SSO-AB was created by multiple panels of experts, it is an attractive gold standard.

It is important to note that the SSO-AB is not an EBM effort. Although randomized controlled trials are important, many users benefit from articles (e.g., reviews, case reports) that are usually excluded from EBM collections. The most prevalent MEDLINE publication types in the SSO-AB are randomized controlled trials (16.4%) and reviews (14.7%). However, 53% of articles are labeled with "Journal article" or other descriptor (please see Appendix, available as a *JAMIA* online supplement at http://www.jamia.org, Table A1).

Algorithms (Retrieval versus Ranking)

In this study, we discuss two types of algorithms: algorithms that retrieve articles from PubMed (labeled "retrieval" in Table 1) and algorithms that rank article, which were previously retrieved from PubMed (labeled "ranking" in Table 1). The experimental design is shown graphically in Figure 1.

In the case of clinical queries and simple PubMed queries, the results are ranked in approximate reverse chronological order. Therefore, we restricted the date range to correspond to the oldest and newest SSO-AB articles (March 1969 through September 2001). Since other algorithms do not rank by date, we did not restrict their date ranges. For "ranking" algorithms, we generated a preliminary result set using simple PubMed queries (please see Appendix 1, available as a *JAMIA* online supplement at http://www.jamia.org, Tables A2 and A3) without date restriction. Whereas each "retrieval" algorithm ranks a potentially distinct result set, "ranking" algorithms all operate on exactly the same result set.

Zettair (TF*IDF)

Zettair is a public domain search engine based on the vector space model that uses TF*IDF‡ weighting. TF*IDF has been found to work well in a variety of applications. A detailed review of the vector space model and associated term

‡Results are actually reported in reverse order of entry into the PubMed database. In some cases, this is not exactly the same as reverse chronological order.
Term frequency (TF) = number of times that a specific term occurs in a given document. Therefore, TF will be large when the term occurs multiple times in a document. Document frequency = Number of number of documents which contain the specific term. Therefore, the inverse document frequency (IDF) will be large for terms which occur in a small number of documents.

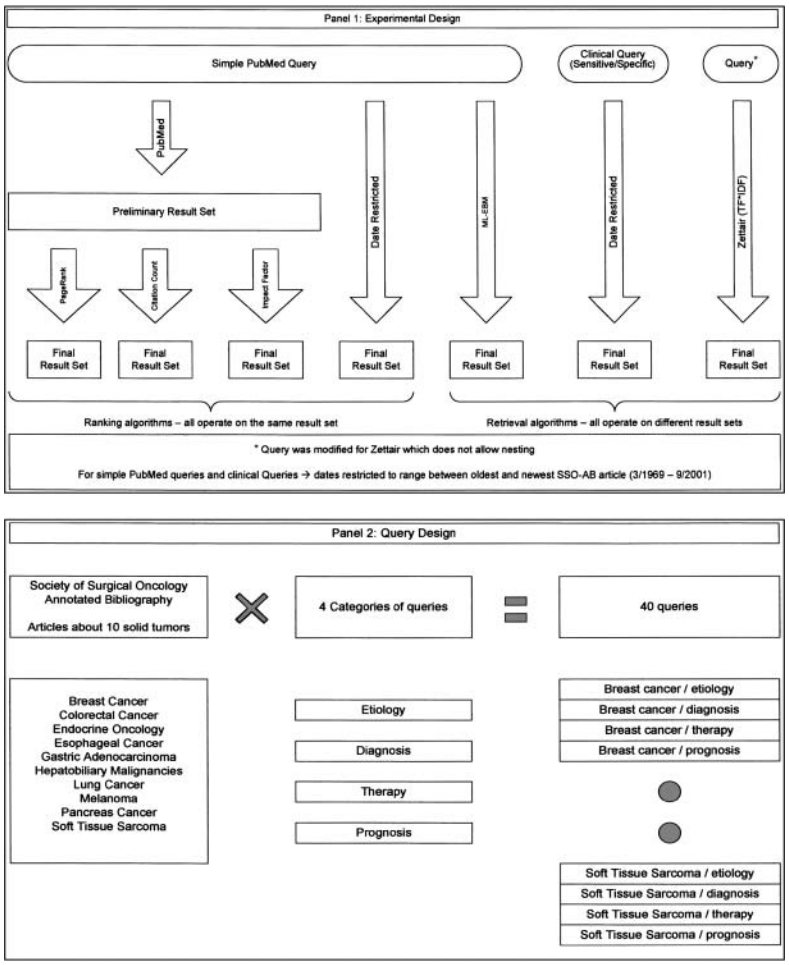

Figure 1. Experimental and query design.

weighting strategies is beyond the scope of this paper. The interested reader is referred to Baeza-Yates and Ribeiro-Neto[22] (specifically section 2.5.3).

Journal Impact Factor (JIF)

The journal impact factor is the number of citations in a given year to a journal's articles that were published over the past two years, divided by the number of articles published over the past two years.[23,24]

PageRank Implementation

We implemented an iterative version of PageRank based on the algorithm described by Brin and Page.[18] We draw an analogy between Web pages and MEDLINE articles. Similarly, Web links are analogous to citations. Therefore, the combination of MEDLINE with the SCI® is a directed graph, similar to the WWW.

Our implementation of PageRank depends on a bidirectional mapping of MEDLINE articles to citations in the SCI® and vice versa. Similarly, citation count (CC) depends on a unidirectional association between MEDLINE articles and citations in the SCI®. The SCI® represents articles as a hash consisting of reference information such as author name, journal, volume, and page numbers expressed as short strings. We therefore construct such a hash and attempt to match. If this succeeds, we return the match. If not, then for unidirectional mapping (CC), we attempt a substring hash match and return a match if there are two or fewer mismatches. Hash substring matching for bidirectional mapping is computationally expensive and is not performed. (Please see Appendix 1, available as a *JAMIA* online supplement at http://www.jamia.org, for pseudo-code of our PageRank implementation.)

Machine Learning–Evidence-based Medicine (ML-EBM)

Machine learning–evidence-based medicine uses a content/quality model built using polynomial support vector machines trained with the *ACP-JC* gold standard. The resulting model is applied to MEDLINE articles and assigns a score to each article that represents the degree of inclusion/exclusion in the gold standard.[25] We used the implementation of these models found in EBMSearch© version 0.3 available at http://ebmsearch.org. In a recent evaluation, the polynomial SVMs performed better than other machine learning methods and clinical query templates. However, ML models for the diagnosis and prognosis categories were preliminary (compared to etiology and therapy) due to reduced sample size. A full description of the development and evaluation of the models is available in Aphinyanaphongs et al.[17,25] ML-EBM is considered a "retrieval" algorithm because only articles with abstracts are ranked. Therefore, the result set is not exactly the same as that returned by PubMed simple queries.

Queries

Evidence-based medicine traditionally distinguishes between articles that address etiology, diagnosis, therapy, and prognosis. This distinction is reflected in the seminal work of Haynes et al., which led to the development of PubMed clinical queries.[26] Similarly, ML-EBM models were developed using the *ACP-JC* gold standard, which relies on the same four categories.[17]

Therefore, we divided the articles referenced in the SSO-AB into four nondisjoint sets reflecting those articles that were judged by NLM indexers to fall under "therapy" [subheading], "etiology" [subheading], "diagnosis" [subheading], and "prognosis" [MeSH terms]. We chose the above terms because they gave a reasonable separation between the categories and reflected human judgment. "Prognosis" appears in multiple places in the MeSH hierarchy and "prognosis" [subheading] is actually under "diagnosis" [subheading]. Therefore, we chose to use the MeSH term rather than the subheading. We did not attempt to force the SSO-AB articles into a single category because some articles address more than one category. For example, it is common for articles to describe how to diagnose and treat a condition.

For each of the ten solid tumors addressed by the SSO-AB, we executed four queries (etiology, diagnosis, therapy, and prognosis). The simple PubMed queries were intended to reflect general information needs, such as "How do you treat breast cancer?" Therefore, result sets were very large. Each algorithm executed a total of 40 queries (Fig. 1), and we report averages for the 40 queries. Each of the ranking algorithm queries used the PubMed search engine to retrieve a preliminary result set based on PubMed simple queries (please see Appendix 1, available as a *JAMIA* online supplement at http://www.jamia.org, Tables A2 and A3).

For experiments using the Zettair search engine, which does not have a PubMed interface, we used a complete local copy of the PubMed database current as of April 18, 2005. We used the same article representation as the PubMed "related articles" feature.[27]

An article was considered a "hit" if it was listed in the SSO-AB under the appropriate topic and had the appropriate category designation. For example, if the query was related to [breast cancer, therapy], we counted articles listed in the SSO-AB under the breast cancer topic which are also associated with the "therapy" [subheading] MeSH descriptor.

Citation Lag

To determine the effect of citation lag, we repeated the experiments using a subset of the data. Specifically, we excluded articles and citations that occurred after 2001. Since the SSO-AB was published in late 2001, this was a simulation of expected performance at the time that the SSO-AB was published. Similarly, we compared the 2001 JIFs to the latest 2003 version.

Evaluation and Data Analysis

Traditional information retrieval evaluation measures do not explicitly consider ordering. For example, recall and precision do not differentiate between algorithms that present the relevant results at the beginning or end of the result set.[22] Therefore, we report recall and precision at 10, 20, 50, 100, 500, and 1,000 results. These values correspond to the observation that information needs can be classified into the need to (1) solve a certain problem or make a decision, (2) obtain background information on a topic, and (3) keep up with information on a topic.[28] Alternatively, Wilkinson and Fuller[29] describe four distinct user needs for document collections such as MEDLINE: (1) fact finding, locating a specific item of information; (2) learning, developing an understanding of a topic; (3) gathering, finding material relevant to a topic that may not be explicitly stated; and (4) exploring, where the information need may not be stated or may change as content is viewed.

Clearly, different search strategies are required to satisfy diverse information needs. Those seeking to keep up with the latest literature on a focused topic may be best served by a system that lists results in reverse chronological order, such as PubMed. We reasoned that a user searching for a single good article would be willing to review at most ten to 20 results. Similarly, a user wanting a brief overview would be willing to review at most 50 to 100 results. Finally, a highly motivated user who needs a comprehensive review may be willing to review 500 to 1,000 results. We recognize that few, if any, routine users would be willing to look through 500 to 1,000 results. However, a user working on a systematic review of a topic may be willing to look through a large result set.

Hit Curves

The SSO-AB (457 unique articles) is much smaller than the result sets that, for simple PubMed queries, range between 4,690 and 124,884 articles. The problem of detection has recently been addressed by statisticians using the concept of a hit curve.[30] In our case, the hit curve h(n) is the number of important articles among the first n results. Hit curves provide an intuitive representation of algorithm performance for a given query or averaged over a number of different queries. If there are x important articles, then the ideal hit curve will be a straight line with a slope of 1, for $1 < n < x - 1$, which becomes horizontal for $n > x$, when all x important articles are retrieved. The ideal search algorithm would produce this kind of hit curve for every query.

Comparing Algorithms Using Hit Curves

We can compare the performance of search algorithms by comparing their hit curves. Intuitively, if we obtain the hit curves resulting from the execution of several search algorithms for the same query, the algorithm corresponding to that hit curve that is "closest" to the ideal hit curve is the best algorithm. In other words, the distance to the ideal hit curve is a measure of performance.

The distance to the ideal hit curve can be measured by several metrics, including maximal vertical separation (distance) between the hit curves, least squares error, and difference in the area under the curve. In all three cases, the smaller the difference, the better the algorithm. To account for the fact that different queries produce different hit curves for the same algorithm, we use the following procedure:

1. For a given query q, normalize the hit curves produced by the ideal algorithm and by the algorithm under consideration with respect to the maximum number of important articles corresponding to q.
2. Compute the distance measures described above: maximum distance, least squares error, and area under the curve, with respect to the normalized curves developed in step 1.
3. Repeat steps 1 and 2 above over a variety of input queries and perform appropriate statistical analyses, such as analysis of variance (ANOVA), on the resulting observations.

Recall and Precision

A more traditional way to compare information retrieval strategies is to compare the mean recall and precision at various result-set sizes. We report the recall for ranking algorithms: simple queries, impact factor, CC, and PageRank. The denominator for the recall calculations was the number of articles in the SSO-AB section on this tumor that fell into the appropriate EBM category (etiology, diagnosis, therapy, prognosis). For the retrieval algorithms, calculations of recall and precision are approximate because both relevance and importance must be considered and are difficult to distinguish. Therefore, an algorithm that does not distinguish between etiology and diagnosis might achieve an unfairly high or even invalid (>100%) recall.

Results

Number of Important Articles Compared to Size of the Result Set

The average size of the result set was 67,145 (range, 4,690–124,884 articles). There were a total of 1,001 nonunique important articles identified by simple PubMed queries. Overlap between categories was allowed. On average, there were 25 important articles per query, which represented 0.15% of the result set (range, 0.07%–1.56%). (See details in Appendix 1, available as a *JAMIA* online supplement at http://www.jamia.org, Table A4.)

Comparison between Algorithms

Figure 2 shows the hit curves for the algorithms tested. Table 3 shows the statistical comparison of hit curves. The citation-based algorithms (CC and PageRank) were more effective at identifying important articles compared to noncitation-based algorithms. The performance of JIF was intermediate between true citation-based algorithms and noncitation-based algorithms.

We found significant differences in favor of the citation-based algorithms ($p < 0.001$) compared to the best noncitation-based algorithm (clinical queries, specific) using hit curve analyses or average precision. When we compared CC to PageRank, the results were not as clear cut. Area under the curve and least squares measures were significant with $p < 0.002$, but maximum distance was not significant ($p > 0.5$). Analysis of the raw data reveals that because of the small number of hits relative to the size of the result set, the ideal hit curve peaks very early, an average of 25 articles until plateau. The maximum distance to the ideal hit curve for many queries is 1 (maximum possible for a normalized curve). Therefore, maximum distance is less sensitive than least squares or area under the curve for our data set.

Table 3 shows the number of important articles retrieved and the mean interpolated average precision[31] calculated over the entire result set. For completeness, we also analyzed a date-limited version of ML-EBM where only articles published within the date range of the SSO-AB were ranked. The difference between date-limited and nondate-limited performance was minor (mean interpolated average precision 0.009 versus 0.006 for nondate-limited, standard deviation 0.016). Table 4 shows recall and precision for ranking algorithms. Retrieval algorithms were excluded from Table 4 because recall-precision calculations were approximate, but recall-precision curves were calculated using approximate data.

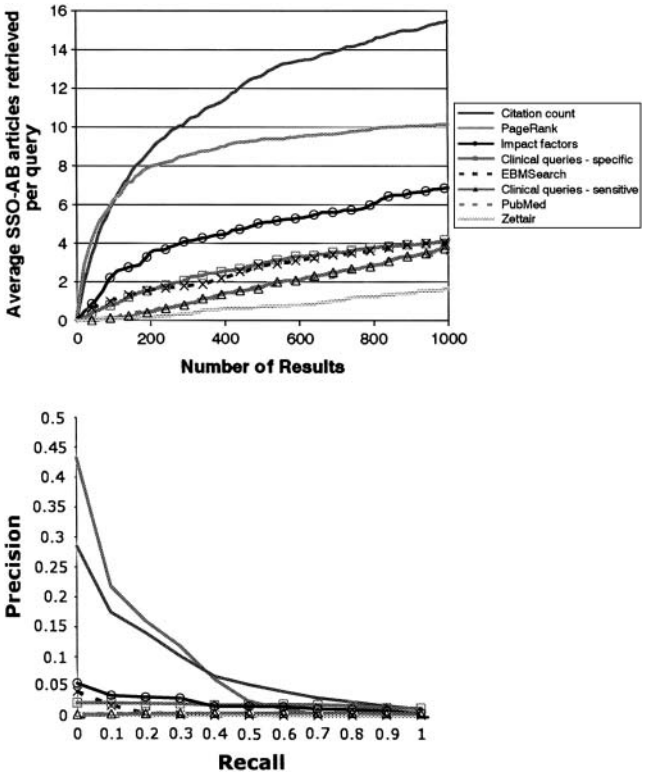

Figure 2. Hit and recall-precision curves.

Figure 2 shows the recall-precision curves for all algorithms. In this analysis, the total number of important articles is assumed to be the number of SSO-AB articles retrieved by the algorithm in the entire result set. Although recall-precision curves incorporate ranking less directly than hit curves, the results appear qualitatively similar.

If a user were to review 100 results presented by a sensitive clinical query, on average, he or she would encounter approximately one important article. In contrast, if he or she were to review the same result list ranked using the JIF, he or she

Table 2 ■ Comparison of Citation Count and PageRank to Clinical Queries (Specific)

Algorithm	Mean Interpolated Average Precision	Mean Deviations from the Ideal Hit Curve Average of 40 Queries		
		Maximum Distance	Δ Area under the Curve	Least Squares
Clinical queries (specific)	0.016	0.99	892.8	816.8
Citation count	0.086	0.91*	603.0*	402.7*
PageRank	0.093	0.88*	700.3*	512.2*
Citation count (2001 data)	0.073*	0.93*	660.4*	475.2*
PageRank (2001 data)	0.045†	0.94*	810.6*	676.1*

*Significant differences with the clinical queries (specific) (one-sided t-test with $p < 0.001$).
†$p < 0.002$.

Table 3 ■ Number of Important Articles Retrieved at 10, 20, 50, 100, 500, and 1,000 Results (Averaged over 40 Queries per Algorithm)

	No. of Results						Interpolated Average Precision[a,b]	
Algorithm	10	20	50	100	500	1,000	Mean	SD
Retrieval								
CQ (sensitive)	0.00	0.00	0.00	0.13	1.75	3.73	0.006	0.006
CQ (specific)	0.08	0.15	0.48	0.80	2.90	4.18	0.016	0.034
ML-EBM	0.20	0.30	0.50	0.85	1.95	3.43	0.006	0.011
Zettair	0.00	0.03	0.05	0.08	0.73	1.65	0.002	0.002
Ranking								
Citation count	1.03	1.98	3.85	6.15	12.63	15.50	0.086	0.083
Impact factor	0.20	0.48	1.00	2.38	5.05	6.85	0.021	0.029
PageRank	1.60	2.90	4.63	6.25	9.38	10.15	0.093	0.078
PubMed (DL)	0.00	0.00	0.00	0.15	1.78	3.48	0.005	0.006
p (ANOVA)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	

CQ = Clinical Queries; DL = Date Limited (3/1969–9/2001); ANOVA = Analysis of Variance; ML-EBM = Machine Learning—Evidence-Based Medicine.

[a]Interpolated average precision = 0 for queries with no hits.

[b]Interpolated average precision calculated at 11 recall levels (0–100%) as described in ref. 22.

would encounter two important articles. If he or she were to use a simple PubMed query ranked using PageRank or CC, he or she would encounter six important articles.

Since performing multiple comparisons increases the risk of type 1 error, we compared only the best noncitation algorithm (clinical queries, specific) to the citation algorithms (PageRank and CC). The citation algorithms were significantly closer to the ideal hit curve by all three measures (maximal distance between hit curves, least squares distance and area under the curve; $p < 0.001$). Citation-based algorithms were more effective at identifying important articles compared to noncitation-based algorithms.

Figure 3 shows the effects of citation lag. In these experiments, we determined the performance of citation-based algorithms (CC and PageRank) using 2001 data and compared it to performance using 2005 data. We found that PageRank was more affected by citation lag than CC. However, using late 2001 data, both algorithms still performed better than noncitation-based algorithms, as shown in Table 2. Similarly, we used 2003 JIFs in the previous experiments and repeated the experiments using 2001 data. The results were practically indistinguishable and the difference was not statistically significant ($p = 0.632$).

Discussion

To our knowledge, this is the first comparative evaluation of a variety of algorithms for identifying important articles in the biomedical literature. We found citation-based algorithms to be more effective than noncitation-based algorithms at identifying important articles. CC was generally more effective than PageRank, a more complex algorithm. Issuing a broad query that returns a large number of results ordered by CC was more effective than quality filtering.

PageRank versus CC

Although the hit curve for PageRank is steep in the beginning, it plateaus lower than CC. We were surprised to find that CC performed better than PageRank, which has been so successful on the WWW. A possible explanation is that PageRank requires a bidirectional mapping of PubMed onto the SCI® and may be more sensitive to mapping errors. CC requires only that the mapping algorithm return the number of citing articles. In contrast, PageRank requires a more sophisticated mapping that returns the identity of the citing articles and the number of citations to each citing article.

Mapping between MEDLINE and the SCI is difficult because article representations are not compatible. For example, panel-authored articles have author lists in the SCI, but not in MEDLINE. Further, there are multiple data entry errors that make simple string matching inadequate. A preliminary evaluation of our mapping methods showed that the function used for CC was over 79% accurate, compared to 70% for the PageRank mapping function. Therefore, mapping errors may have affected PageRank more than CC.

The effects of better mapping are difficult to predict. In other work, we explored the relationship between citation database completeness and performance. Specifically, we randomly removed citations and evaluated the ability of PageRank and CC to identify SSO-AB articles. We found that performance did not degrade significantly until over 95% of citations were removed.[32] If we assume that mapping errors are randomly distributed, then mapping errors should not have had a significant impact on relative results. In other words, with perfect mapping, results should remain similar.

Table 4 ■ Recall and Precision for Ranking Algorithms at 10, 20, 50, 100, 500, and 1,000 Results (Averaged over 40 Queries per Algorithm)

	No. of Results					
Algorithm	10	20	50	100	500	1,000
Citation count						
Recall	4.67%	8.88%	17.28%	26.43%	52.29%	62.68%
Precision	10.25%	9.88%	7.70%	6.15%	2.53%	1.55%
Impact factor						
Recall	1.20%	2.76%	5.45%	11.89%	23.65%	31.04%
Precision	2.00%	2.38%	2.00%	2.38%	1.01%	0.69%
PageRank						
Recall	7.77%	13.01%	19.89%	26.29%	37.30%	40.94%
Precision	16.00%	14.50%	9.25%	6.25%	1.88%	1.02%
PubMed*						
Recall	0.00%	0.00%	0.00%	0.93%	9.50%	17.40%
Precision	0.00%	0.00%	0.00%	0.15%	0.36%	0.35%

*Dates limited to the date range of the Society of Surgical Oncology.

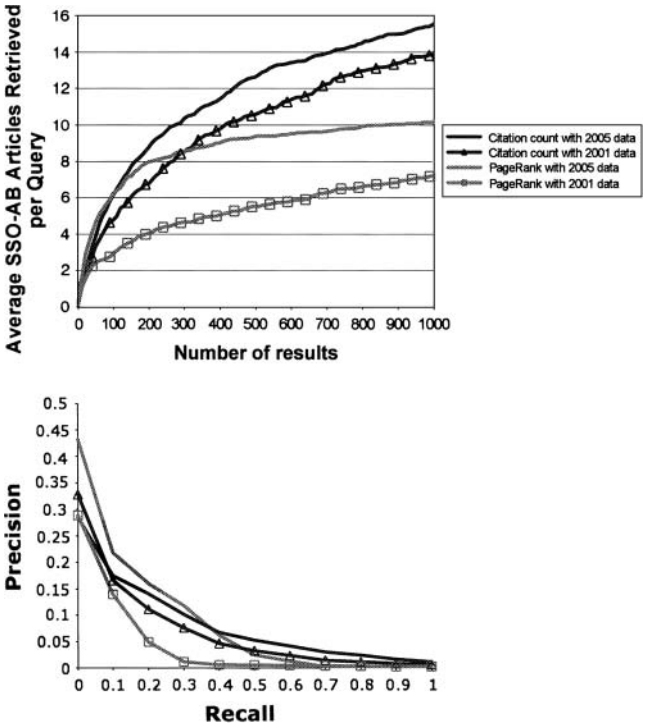

Figure 3. Comparison of PageRank and citation count using 2001 versus 2005 data (hit curves and recall-precision curves).

Kleinberg[19] argues that simply counting the number of incoming links is not an effective ranking strategy for WWW search. He observes that popular sites such as http://www.yahoo.com will be ranked highly whenever such a site is returned, regardless of whether it is actually an authoritative site on that topic. For example, Yahoo!® may mention Volkswagen automobiles and has many incoming links. However, the Volkswagen company site is likely a better source of information about Volkswagen automobiles than Yahoo!®. Kleinberg concludes that more sophisticated algorithms are required to identify authoritative sites.

In contrast, we found CC to perform at least as well as PageRank for identifying important articles (i.e., those listed in the SSO-AB). A possible explanation is that citation patterns in the biomedical literature differ from WWW linking patterns. Although there may be other reasons, authors often cite to provide background or to support a method or statement.[33] Therefore, unlike the WWW, authoritative articles are most likely to be cited. Methodology articles such as those describing techniques for statistical analysis or laboratory methods may be exceptions. However, these are the minority of clinical citations.

We found that citation lag affected PageRank more than CC. CC requires only one degree of citation. In other words, to determine the CC of article A, you simply count the number of articles that cite A. In contrast, the PageRank of article A depends on knowing the number of citations to the articles that cite A, and so on recursively. Therefore, PageRank depends on citing articles to be cited, which takes more time.

Journal Impact Factor

Ranking articles using JIF is appealing because it is very simple to implement. Instead of having to maintain a database of citations, one only has to perform a table lookup. Further, JIFs can be calculated for journals not currently tracked by Journal Citation Reports.[34] Unfortunately, JIF does not perform as well as CC or PageRank. This may be due to multiple letters, commentaries, and other nonstudy publications that, if published in a high-impact journal, are ranked highly. In future work, we will explore combining JIF with publication information. For example, a randomized clinical trial could be ranked above a letter that appears in the same journal.

Limitations and Comparison to Previous Studies

There are no undisputable gold standards in information retrieval. One can always question whether a particular article is relevant to a given query. Similarly, importance is inherently subjective. We chose the SSO-AB because it was compiled by panels of experts choosing, in their own words, "important literature on common problems in surgical oncology." The SSO-AB predates this project and is not related to our effort. Therefore, information retrieval issues did not influence the gold standard.

If an article is listed in the SSO-AB, then we can conclude that it is important. However, if an article is not listed in the SSO-AB, we cannot necessarily conclude that it is not important. Therefore, it is conceivable that algorithms that do not score SSO-AB articles highly may be identifying other important articles. This is true for any reasonably sized general purpose collection of important literature on a broad topic.

In this study, we used the second edition of the SSO-AB published in 2001. The first edition, published one year earlier, may have influenced citation patterns. Since 310 of 457 (68%) articles are shared between versions, our results could have been affected. This effect is difficult to quantify because articles that were found in the first version are likely to be older and therefore, all other factors being equal, more highly cited than newer articles.

We found that ML-EBM, which is based on machine learning using polynomial SVMs, was approximately as effective as sensitive clinical queries that were developed using human effort. In this, our findings are similar to the previously published evaluation.[17] Again, we note that ML methods trained using the SSO-AB rather than another article collection (*ACP-JC*) would likely perform better. Further, ML methods incorporate methodological quality and topic and these are difficult to separate. Similarly, clinical queries were developed to retrieve articles that meet the *ACP-JC* inclusion criteria.[15] Although clinical queries are not intended to be domain specific, it is possible that their performance would be better in the domain of internal medicine compared to surgical oncology. Therefore, performance of ML-EBM and clinical queries was probably compromised by switching topics.

For the above reasons, we do not consider our evaluation to be a competition between algorithms. Instead, we believe alternative approaches to be complementary. For example, a user can retrieve high-quality articles using clinical query templates and then rank them using CC.

Citation-based algorithms such as CC and PageRank are context free. In other words, they do not depend on the topic (e.g., internal medicine versus surgical oncology), particular gold standard (e.g., *ACP-JC* versus SSO-AB), or specific query. Therefore, we can expect CC and PageRank to be equally effective in a variety of fields, as long as citation

patterns are similar. Further, they cleanly separate importance from relevance and quality.

Citation tracing, starting with a seed article and then following its references to retrieve the relevant past literature, is a common strategy in manual review. Although we do not know precisely how articles were chosen by experts for inclusion in the SSO-AB, it is possible that citation tracing may have been used. If true, citation tracing may partially explain why citation-based algorithms correlated with expert opinion in our study.

In our study, all algorithms performed poorly compared to previous evaluations. Recall and precision for MEDLINE information retrieval vary widely, but generally fall in the 10% to 75% range depending on user, available tools, and specific task.[8] In our evaluation, even the best algorithms had a precision of less than 2% at 1,000 articles. However, the task in our evaluation was substantially different. Specifically, we used broad queries with result sets that were over 1,000 times larger than other studies (see Hersh and Hickam,[8] Table 2, for example). Further, many previous studies evaluated systems with respect to their ability to identify relevant articles. Our evaluation included both relevance and importance. The low precision at 1,000 articles does not reflect an unusual information need. Rather, it reflects the small size of an annotated bibliography compared to large result sets intended to reflect general information needs. Therefore, it is not appropriate to compare our quantitative results directly to those of previous studies.

Future Work

We evaluated algorithms with respect to their ability to retrieve articles from a gold standard collection (SSO-AB). However, most users do not evaluate their searches against a gold standard collection. Instead, they determine whether the search satisfies their information need. The ability to retrieve articles from a gold standard collection does not necessarily mean that the algorithm can satisfy actual information needs. An evaluation that uses actual users with real information needs could address this question but is more difficult to conduct and may not generalize since information needs are inherently specific to a given user and situation.

Analysis of actual usage provides the best evaluation. Therefore, we are collecting click-through data using our MEDLINE interface, which is available to users at our institution. A click-through occurs when the user clicks on a search result to obtain more information, such as the abstract. The underlying assumption is that a click-through is an expression of interest in the article and users preferentially click on useful articles. Click-through data will allow us to compare algorithms with respect to the interest that their results attract from users who issue specific queries. This approach was suggested by Joachims,[35] but to our knowledge has not yet been applied in the biomedical domain.

Conclusion

Our results suggest that citation-based algorithms can identify important articles from large result sets more effectively than algorithms based on vector space article representations (Zettair), Boolean queries (clinical query templates), and models based on ML (ML-EBM). In addition, we found that mapping the SCI® onto PubMed/MEDLINE and vice versa is a difficult task that may limit the utility of PageRank, which requires a sophisticated mapping strategy. Further research is needed to determine whether our results generalize to other domains, other gold standards, and real user information needs. Citation-based algorithms may complement other strategies, such as ML and clinical query templates. Information retrieval systems of the future will likely leverage multiple strategies, including domain knowledge and citation analysis, to meet users' information needs.

References ■

1. Zipser J. MEDLINE to PubMed and beyond. Available from: http://www.nlm.nih.gov/bsd/historypresentation.html/. Accessed 2005 Mar 29.
2. PubMed. Available from: http://www.ncbi.nlm.nih.gov/pubmed/. Accessed 2005 Mar 29.
3. NLM. MEDLINE Fact Sheet. WWW. September 18, 2002. Available from: http://www.nlm.nih.gov/pubs/factsheets/medline.html/. Accessed 2003 Apr 30.
4. Wilson SR, Starr-Schneidkraut N, Cooper MD. Use of the critical incident technique to evaluate the impact of MEDLINE. Final Report. Palo Alto, CA, 1989.
5. Marshall JG. The impact of information provided by the hospital libraries in the Rochester area: Rochester Regional Library Council, 1991.
6. King DN. The contribution of hospital library services to patient care. Bull Med Library Assoc. 1987;75:291–301.
7. Michigan Health Sciences Libraries Association Research committee report. Ann Arbor, MI: University of Michigan, 1992.
8. Hersh W, Hickam D. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. JAMA. 1998;280:1347–52.
9. NLM. NLM resource lists and bibliographies. April 1, 2003. Available from: http://www.nlm.nih.gov/pubs/resources.html/. Accessed 2003 May 2.
10. Hersh WR, Detmer WM, Frisse ME. Information-retrieval systems. In: Shortliffe EH, Perreault LE, Wiederhold G, Fagan LM, editors. Medical informatics: computer applications in health care and biomedicine. 2nd ed. New York: Springer-Verlag, 2000. p. 539–72.
11. PubMed Clinical Queries. Available from: http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.shtml/. Accessed 2005 Apr 27.
12. Opthof T. Sense and nonsense about the impact factor. Cardiovasc Res. 1997;33:1–7.
13. Borodin A, Roberts GO, Rosenthal JS, Tsaparas P. Link analysis ranking: algorithms, theory and experiments. ACM Trans Internet Technol. 2005;5:231–97.
14. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. J Am Med Inform Assoc. 2002;9:653–8.
15. Haynes RB, Wilczynski N. Finding the gold in MEDLINE: clinical queries. ACP J Club. 2005;142:A8–9.
16. NLM. PubMed clinical queries table. Available from: http://www.ncbi.nlm.nih.gov/entrez/query/static/clinicaltable.html/. Accessed 2005 Mar 31.
17. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc. 2005;12:207–16.
18. Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Paper presented at WWW7/Computer Networks, April 14–18, 1998, Brisbane, Australia, 30(1–7), pp. 107–17.
19. Kleinberg J. Authoritative sources in a hyperlinked environment. Paper presented at Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, January 25–27, 1998, San Francisco, CA. Available at http://www.cs.cornell.edu/home/kleinber/auth.ps.

20. Search engines take a quantum leap: 19 out of 20 now use link popularity to determine relevancy. Available from: http://www.webseed.com/page1007.html/. Accessed 2001 Mar 14.
21. Hersh WR, Buckley C, Leone TJ, Hickam D. OHSUMED: An interactive retrieval evaluation and new large test collection for research. Paper presented at SIGIR'94, 1994, Dublin, Ireland.
22. Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. New York: Addison Wesley and ACM Press, 1999.
23. Science Citation Index. Journal citation reports. Philadelphia: Thomson-ISI, 2003.
24. Garfield E. Journal impact factor: a brief review. CMAJ. 1999;161: 979–80.
25. Aphinyanaphongs Y, Aliferis CF. Learning Boolean queries for article quality filtering. Paper presented at Medinfo, 2004, San Francisco, CA.
26. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc. 1994;1:447–58.
27. Computation of related articles. February 6, 2003. Available from: http://www.ncbi.nlm.nih.gov/entrez/query/static/computation.html/. Accessed 2005 Apr 28.
28. Lancaster F, Warner A. Information retrieval today. Arlington, VA: Information Resources Press, 1993.
29. Wilkinson R, Fuller M. Integration of information retrieval and hypertext via structure. In: Agosti M, Smeaton A, editors. Information retrieval and hypertext. Norwell, MA: Kluwer, 1996. pp. 257–71.
30. Zhu M, Chipman HA, Su W. An adaptive method for statistical detection with applications to drug discovery. Paper presented at Joint Statistical Meetings, Biopharmaceutical Section, 2003, San Francisco.
31. Manning CD, Schutze H. Foundations of statistical natural language processing. Cambridge, MA: MIT Press, 1999.
32. Herskovic JR, Bernstam EV. Using incomplete citation data for MEDLINE results ranking. Paper presented at American Medical Informatics Association Fall Symposium, 2005, Washington, DC.
33. Garfield E. Can citation indexing be automated? Vol 1. Philadelphia: ISI Press, 1977.
34. Stegmann J. How to evaluate journal impact factors. Nature. 1997;390:550.
35. Joachims T. Evaluating retrieval performance using clickthrough data. Paper presented at SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval, 2002, Tampere, Finland.
36. Williams H. The Zettair Search Engine. August 10, 2004. Available from: http://www.seg.rmit.edu.au/zettair/. Accessed 2005 Apr 27.
37. Science Citation Index Expanded (2004 Q4 update). Vol 2004. Philadelphia: Thomson-ISI.
38. Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: bringing order to the web. Stanford digital libraries electronic source; 1998. Available at http://dbpubs.stanford.edu/pub/1999-66.