Journal of Biomedical Informatics 42 (2009) 678–684

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Sequential result refinement for searching the biomedical literature ☆

L.Y. Tanaka [a,b,*], J.R. Herskovic [a], M.S. Iyengar [a], E.V. Bernstam [a,c,*]

[a] *School of Health Information Sciences, The University of Texas Health Science Center at Houston, 7000 Fannin Street, Suite 600, Houston, TX 77030, USA*

[b] *Kapiolani Medical Center for Women & Children, University of Hawaii, John A. Burns School of Medicine, Department of Pediatrics, 1319 Punahou Street 7th Floor, Honolulu, HI 96826, USA*

[c] *Department of Internal Medicine, The University of Texas Medical School at Houston, 6431 Fannin Street, MSB 1.150, Houston, TX 77030, USA*

ARTICLE INFO

Article history:
Received 14 September 2008
Available online 9 March 2009

Keywords:
Algorithms
Bibliometrics
Information storage and retrieval/methods
MEDLINE

ABSTRACT

Information overload is a problem for users of MEDLINE, the database of biomedical literature that indexes over 17 million articles. Various techniques have been developed to retrieve high quality or important articles. Some techniques rely on using the number of citations as a measurement of an article's importance. Unfortunately, citation information is proprietary, expensive, and suffers from "citation lag." MEDLINE users have a variety of information needs. Although some users require high recall, many users are looking for a "few good articles" on a topic. For these users, precision is more important than recall. We present and evaluate a method for identifying articles likely to be highly cited by using information available at the time of listing in MEDLINE. The method uses a score based on Medical Subject Headings (MeSH) terms, journal impact factor (JIF), and number of authors. This method can filter large MEDLINE result sets (>1000 articles) returned by actual user queries to produce small, highly cited result sets.

1. Introduction

MEDLINE is the world's largest bibliographic database of biomedical science. MEDLINE indexes over 17 million articles, and grows by over 2500 records every day [1]. This volume of information makes MEDLINE's simple searches ineffective for many information needs and demands that users know how to use advanced features. Many MEDLINE users, however, do not use the system proficiently. Only 20% of queries use boolean operators. The average size of a query's result set is over 14,000 articles [2]; too many to review. In fact, most users never go beyond the first page of results [3]. PubMed (MEDLINEs free web interface) users by default only see the newest 20 articles for any query; missing out on the vast majority of potentially useful results. Whether these 20 articles actually fulfill the users information need is unknown.

In practice, different users have different information needs. For example, medical students may need information on patient care topics, and faculty may search for research in their field or for conference proceedings [4]. We traditionally evaluate search systems through the concept of "relevance." In biomedical information retrieval, a relevant article is an article that satisfies the information need of the user. A good retrieval strategy should produce all of the relevant documents (its recall should be high) and only relevant documents (its precision should be high), answering the user's information need. Unfortunately, given the size of literature databases, even high recall, high precision strategies can produce very large result sets [5]. For example, using PubMed's clinical query filters to search for diagnostic information on "breast cancer" with a "narrow, specific" (i.e., high-precision, lower recall) filter still returns 4220 results [6].

A different, complementary approach to handle the overload is ranking PubMed results using citation information. We define an article's importance as its influence (or future influence) in its field of study. Importance is difficult to measure directly, but it can be operationalized through citation analysis. Highly cited articles affect their field more than articles that are never cited. Results from searches can thus be processed and ranked by the number of citations each article receives effectively constructing an importance-ranked list of articles [5].

☆ This work was supported by a training fellowship from the Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia (NLM Grant 5T15LM07093). This paper resulted from a presentation at the National Library of Medicine Training Conference held at Stanford University from July 24–25, 2007. Thanks is given to Marie McVeigh at Thomson Scientific for providing valuable assistance with access to the Science Citation Index, and to Devika Subramanian for comments and advice.

* Corresponding authors. Addresses: Kapiolani Medical Center for Women & Children, University of Hawaii, John A. Burns School of Medicine, Department of Pediatrics, 1319 Punahou Street 7th Floor, Honolulu, HI 96826, USA (L.Y. Tanaka); School of Health Information Sciences, The University of Texas Health Science Center at Houston, 7000 Fannin Street, Suite 600, Houston, TX 77030, USA (E.V. Bernstam).

E-mail addresses: lent@hawaii.edu (L.Y. Tanaka), Jorge.R.Herskovic@uth.tmc.edu (J.R. Herskovic), M.Sriram.Iyengar@uth.tmc.edu (M.S. Iyengar), Elmer.V.Bernstam@uth.tmc.edu (E.V. Bernstam).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.02.009

However, citation analysis algorithms require citation databases. Papers are not cited immediately after publication, but after being read by authors who then write and publish their own papers. Citation information therefore appears slowly. This citation lag affects all citation analysis algorithms and makes them less effective. In addition, high-quality citation databases, like the Science Citation Index (SCI) (The Thomson Corporation, Stamford, CT), are expensive. Therefore, we were motivated to develop a method to identify a subset of articles likely to be highly cited using only information freely available at the time of an article's listing in MEDLINE.

2. Background

2.1. Approaches to handling information overload

Previous approaches to handling information overload include summarization, query assistance, and machine learning. There are also subscription services that provide summaries of new research such as UpToDate [7]. These services require domain experts to conduct rigorous reviews and write summary documents. Query templates or Support Vector Machines (SVMs) have been used to select high quality articles by methodological criteria (the evidence-based medicine approach) [8,9]. These approaches still return a large number of results for general queries. For example, the most restrictive clinical query filter for treatment still returns 4794 articles on "breast cancer." Aphinyanaphongs and colleagues used SVMs trained on a collection of important articles to generate ranking scores in cancer surgery and were able to rank these articles at least as well as citation-based measures [10]. These classifiers are also limited, as they require retraining for each new task.

2.2. Using citations

A citation consists of bibliographic information such as the title, authors, and journal information. It is, essentially, a "document address" [11]. Authors cite papers for a variety of reasons including providing background, indicating related work, and highlighting areas of controversy [11]. The number of times a paper is cited in the literature can and has been used as a proxy measure for importance, identifying the impact of papers [5,12,13]. Most papers receive zero or one citation [14]. Only 1% of articles receive six or more citations [14]. A cited article has an average of 3.2 citations [14].

Citations between scientific papers can be modeled as a network, similar to the way hyperlinks create a network of pages on the Web. We use this model to explore applications of algorithms developed for the Web to a network of papers (pages) and citations (links). The link structure provides valuable information about the relative importance of pages [13]. Algorithms like PageRank leverage link information to prioritize important articles [15,3]. In previous work, we found that different citation analysis algorithms, like simple citation counts or Pagerank were more effective than non-citation-based ranking algorithms at retrieving important surgical oncology articles on PubMed [5]. Citation analysis does not require explicit human evaluation. However, citation data are expensive in proprietary citation databases, and suffer from citation lag. Total citation counts were as effective as ranking algorithms (including PageRank) in a previous study [5]. In this study, we therefore attempt to predict total citation counts.

2.3. Selecting predictors

Q. L. Burrell studied the citation prediction problem. He derived a probability distribution for the number of future citations to an article based on when and how often the article was cited [16]. However, because this formula required keeping track of when an article is cited, it was still susceptible to citation lag. Researchers have also attempted to predict citation to a journal rather than to a specific article within a journal. Garfield observed that citation frequency is a function of "many variables besides scientific merit" [12]. However, he did not define a way to calculate the number of citations to a journal based on these variables [12]. Instead, Garfield studied citations grouped by journal. The journal impact factor (JIF) is a numerical score based on citations to a journal, described as [17]:

$$\mathrm{JIF}(y) = \frac{\#\ \text{citations in year}, y, \text{to articles published in } y-1 \text{ and } y-2}{\#\ \text{articles published in } y-1 \text{ and } y-2}$$

In fact, Garfield believes that publication in a particular journal influences the number of citations an article receives [18]. Therefore JIF is an obvious predictor of the citation count of an article. JIF tends to remain relatively stable as a journal grows, publishing more articles and receiving more citations [19]. JIF alone, however, should not be used to evaluate individual authors or publications [20]. Therefore, we must consider additional predictors to rank individual articles.

Intuitively, we expect well-known authors to publish and be cited more frequently than less-known authors. Van Dalen and Henkens tested this hypothesis within a small field (demography) [21]. They evaluated both the number of citations to papers published by an author as a measure of reputation, and the number of authors of a paper [21]. Unfortunately, MEDLINE does not track authors' citation history. Based on these previous publications, we decided that predictors for citation in MEDLINE would include at least JIF and number of authors of a paper.

A prior study of MEDLINE found less collaboration in biomedicine than in other scientific disciplines such as physics [22]. In biomedicine, research networks resembled top-down, tree-like structures. These networks probably arose from the fact that it is common for lead scientists to direct labs that publish multiple research papers [22]. Collaboration between authors appears to increase the impact of a publication [23]. Like Van Dalen and Henkens, we use the number of authors associated with an article as a proxy measure for collaboration [21].

We identified additional candidate predictors from the literature and brainstorming sessions. The complete list of candidate predictors was: JIF, JIF squared, number of authors, total sum of all authors past publications, average number of all authors past publications, highest number of past publications, lowest number of past publications, product of all authors past publications, and the SCI categories of anesthesiology, critical care medicine, general medicine, oncology, pediatrics, psychiatry, and surgery. We performed a regression analysis to identify which combination of these variables could predict citation counts. Our regression analysis identified JIF and number of authors as the best predictive variables, which was consistent with previous publications in other fields. However, the regression model was not accurate enough to predict citation counts for individual articles. We therefore decided to focus on predicting which articles will be highly cited. To achieve that, we decided to add topical information to our predictor.

Important research topics occur in waves. Kuhn proposed that new theories cause an increase in the number of publications [24]. Similarly in the biomedical literature, ideas that are highly controversial or groundbreaking lead to follow-up articles, opinions, letters, and editorials. Topic trends in the biomedical literature can be tracked using term counts. For example, Citespace identifies research fronts by tracking term frequencies in titles, abstracts, descriptors, and identifiers [25]. In other fields, the Bursty algorithm can categorize and track email topic changes according to word frequency [26].

All articles in MEDLINE are indexed using a standard vocabulary called the Medical Subject Headings (MeSH). We define the popularity of a MeSH term as the number of articles indexed with that term over a year. Since MEDLINE grows faster every year, we normalize the popularity scores by the total number of articles added that year to allow comparison across years. We thus define a scoring function the MeSH Percentage Article Count Total (MPACT) as:

$$f(m,y) = \frac{n(m,y)}{N(y)}$$

$$\begin{aligned} n(m,y) &= \#\ \text{articles published with MeSH} \\ &\qquad \text{major heading } m,\ \text{in year } y \\ N(y) &= \text{total articles published in year } y \end{aligned}$$

In this paper we attempt to identify articles that will be highly cited using the JIF of the journal in which the article appears, the number of authors, and the MPACT score for the MeSH terms associated with the article in MEDLINE.

2.4. Evaluating searches

Evaluation in information retrieval traditionally relies on relevance-based measures: recall and precision. Recall is the number of relevant documents retrieved, divided by the total number of relevant documents. Precision is the proportion of relevant documents in the result set. Calculating recall is difficult in large and dynamic document collections like the Web. As MEDLINE continues growing, recall becomes more and more difficult to calculate, since it becomes impossible to manually identify all relevant documents. Alternate measures such as the precision at rank *n* (p@n) put less emphasis on recall [27,28]. Further, users rarely look beyond the first page of results, rendering recall impractical in the evaluation of large search systems [3].

For this study we assumed that PubMed users are not interested in finding all possibly relevant documents on broad topics, but instead desired a small number of relevant documents. Therefore we focused on high precision as the primary evaluation criterion. To accomplish this, we constructed a set of MEDLINE articles that we correlated to the SCI. After regression analysis of candidate predictors, we developed a filter for search results using simple thresholds for JIF, MPACT, and number of authors to identify articles that were likely to be highly cited in the future. Finally, we evaluated the filtered results using actual queries submitted to PubMed.

3. Materials and methods

3.1. Data set construction

We downloaded MEDLINE records for articles published in 1994 and added to PubMed by 2004 (articles may be added to MEDLINE years after publication). We accessed MEDLINE using Python version 2.5.1 and BioPython version 1.42 [29,30]. Each MEDLINE record contains up to 52 fields including identification numbers, authors, journal name, categorization terms from MeSH, and free text in the article title and abstract [31]. We used author data from the MEDLINE records.

For citation information we used the SCI database published on CD-ROM (The Thomson Corporation, Stamford, CT). We extracted the information from the 1999–2004 discs (inclusive) into PostgreSQL version 8.1.4 [32]. The SCI records were mapped to corresponding MEDLINE records. This was comparable to the data used in previous research on predicting citation [21]. Although SCI records contain topical data, articles are divided into very general categories like “pediatrics” or “surgery.” Our model does not use the SCI topical data, since MEDLINE records contain more detailed topical information (MeSH terms).

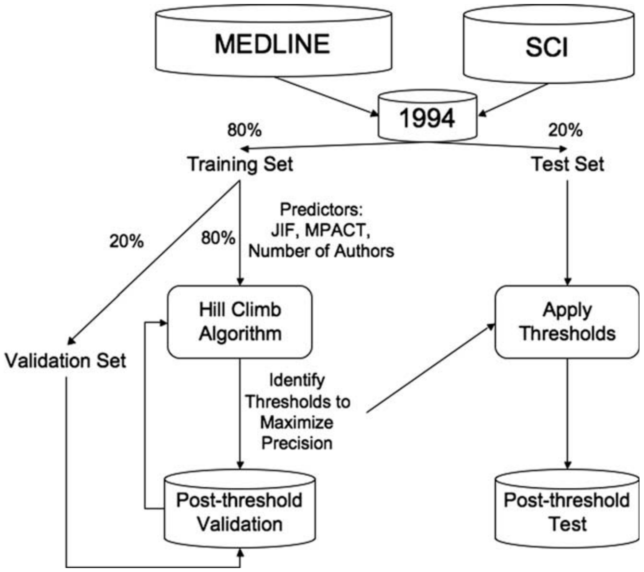

Fig. 1. Diagrammatic representation of Sequential Search Refinement data set creation and algorithm development.

Finally, we obtained JIF information from the ISI Web of Science directly from the Thomson Corporation. We used the most current JIFs available at the time of article publication. For example, JIF information from 1992 would be available at the time of publication for articles published in 1994. Not all journals have a JIF, and we were unable to map approximately 33% of 422,302 PubMed articles in 1994 to the SCI. We thus had JIF and MEDLINE information for 281,873 articles in 1994. We called the set of articles present in both databases the “1994 set.” We created a second set in the same fashion for articles published in 1999. We called it the “1999 set.” We split the 1994 set into a training set (80%) and a test set (20%) containing 225,497 and 56,376 MEDLINE records respectively (see Fig. 1).

3.2. Regression analysis

We conducted a regression analysis of a random subset of 30,782 articles from the 1994 set to determine if the number of citations could be predicted reliably. Candidate predictors included JIF, JIF squared, number of authors, total sum of all authors’ past publications, average number of all authors’ past publications, highest number of past publications, lowest number of past publications, product of all authors’ past publications, and the SCI categories of anesthesiology, critical care medicine, general medicine, oncology, pediatrics, psychiatry, and surgery. We only disambiguated author names to the extent permitted by the MEDLINE record, i.e., by matching the last name and first initial.

Citations to an article accumulate over time as non-negative integers, i.e., count data. Regression methods for the underlying probability distributions included both Poisson and negative binomial [33]. We used SAS/STAT version 9.1 (SAS Institute Inc., Cary, NC) for regression and goodness-of-fit statistics. Significance was set at $p < 0.05$.

3.3. MPACT score

We downloaded MeSH version 2006, which contained 23,883 terms, and counted the occurrence of these terms in articles from 1980 through 2006 as of January 2007 [34]. Previously indexed

MEDLINE records are updated to reflect the current version of MeSH [34]. We determined the average number of edits (173), deletions (46), and additions (596) to MeSH each year over four years. MEDLINE records contain starred MeSH terms called major headings that "reflect the central concepts of an article" [35]. Indexers assigned MeSH terms to each article between one to eight weeks after information was received from the publisher (S. Nelson, personal communication, November 14, 2007). For each MeSH term, we collected a total article count for each year from 1980 to 2006. In order to correct for the accelerating publication rate, we normalized by dividing the total article count by the total number of articles published in the given year. The MPACT score for an article was the sum of the scores for all major headings assigned to the article.

The MPACT score, $g(M,y)$, for an article is:

$$g(M,y) = \sum_{m\in M} f(m,y)$$

M is a set of MeSH major headings for the article, y is the article publication year, and f is the previously described function.

3.4. Sequential result refinement

Since the reported average number of citations for cited articles was 3.2, we arbitrarily chose to label an article cited five or more times as "highly cited" [14]. We conducted a sensitivity analysis on the frequency of articles cited one or more times through ten in the 1994 set (see Fig. 2).

To identify the optimal thresholds for our predictors (JIF, MPACT, and author count), we developed a hill-climbing algorithm that optimized for precision in the 1994 training set. We defined precision as the percentage of highly cited articles in the result set, i.e., the number of articles cited five or more times divided by the number of retrieved articles. We implemented the hill-climbing algorithm as a gradient ascent method on each predictor with a random start location and a decay parameter to scale nearest neighbor values with the termination criteria being no further improvement in precision [36]. The threshold value found for each predictor provided a point to split the data. We filtered the 1994 set using the thresholds to create a "post-threshold" set. We then calculated the percentage of highly cited articles in the post-threshold set. We took the 1994 training set and split it further, 80/20, to obtain a validation set for algorithm development. The validation set was used to optimize the precision by trying various decay parameters. We selected the decay parameter at the point where precision leveled off. As the parameter space for the algorithm was large, global optimality could not be guaranteed.

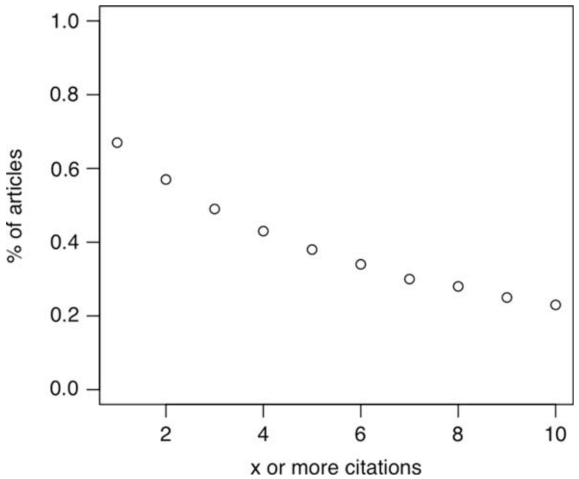

Fig. 2. Sensitivity analysis plot shows various citation amounts for highly cited articles in the 1994 set. We selected an article having five or more citations to be highly cited.

After completing algorithm development, we fixed the thresholds (see Fig. 1). We evaluated the performance of the thresholds with the same evaluation used by the hill climber: by using them to filter the test set and thus construct a "post-threshold" test set, which we compared against the original test set by measuring the average number of citations. We compared the 1994 test set and the 1999 set before and after applying the thresholds using the Chi-square test. Statistical significance was set at $p < 0.05$.

3.5. Evaluation using actual PubMed queries

To evaluate the utility of the thresholds on real searches of the biomedical literature, we created a random sample of queries selected from a large anonymous PubMed query log from October 2005 [37]. Since we were interested in extracting highly cited articles from large result sets, we selected queries that returned more than 1000 results without a date limit.

Our random sample had 426 queries. Our method focused on articles published in 1994, so therefore we matched the PubMed ID field from each result to the pre-threshold and post-threshold sets. This allowed us to determine the percentage of highly cited articles (precision) in both the pre-threshold and post-threshold sets. We compared the results from the queries using the Wilcoxon signed-rank test. We used the publicly available R software package version 2.4.1 for statistics, tabulation, and graphing [38].

4. Results

MEDLINE contained 422,302 articles published in 1994. The majority, 228,796 (54%), had no citations. 281,873 (67%) articles, the 1994 set, were from journals with an available JIF. The mean number of citations in the 1994 set was 8.9 (range: 0–11,732; median: 2). The mean JIF in the 1994 training set was 3.824 (range: 0.024–52.28; median: 2.397). The mean MPACT score was 0.012 (range: 0–0.360; median 0.005). The mean number of authors was 3.8 (range: 0–64; median: 3). There were 86,734 (38.5%) articles with five or more citations (see Fig. 3).

4.1. Regression analysis

We obtained the best fit on the regression analysis using a negative binomial distribution with a deviance/degree of freedom (DF) ratio of 1.0262 and Pearson Chi-square/DF ratio of 1.8327. The statistically significant predictors with a positive coefficient were JIF, number of authors, average number of past publications, and the

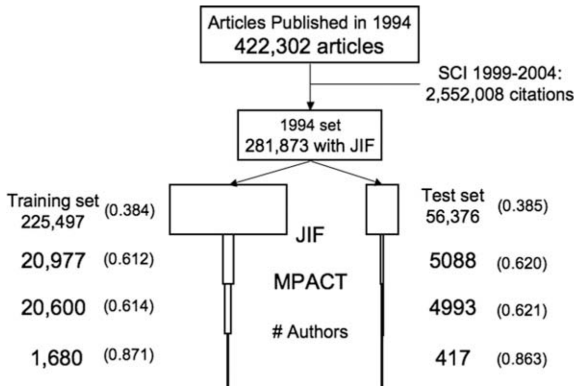

Fig. 3. Sequential Search Refinement reduction in the data set size for articles with five or more citations. Precision values reported in parentheses.

Table 1
Negative binomial regression for predictors.

Factor	Coefficient	*p*-value
JIF	0.3716	<0.0001
JIF2	−0.0116	<0.0001
Number of authors	0.2114	<0.0001
Total number of past publications	−0.0004	<0.0001
Average number of past publications	0.0029	<0.0001
Maximum number of past publications	−0.0002	0.0193
Minimum number of past publications	−0.0019	<0.0001
Product of number of past publications	0	0.0576
SCI category		
Anesthesiology	−0.1931	<0.0001
Critical care	−0.0391	0.5279
Medicine, general	−0.6364	<0.0001
Oncology	0.1986	<0.0001
Pediatrics	0.0582	0.1021
Psychiatry	0.1087	0.0039
Surgery	–	–

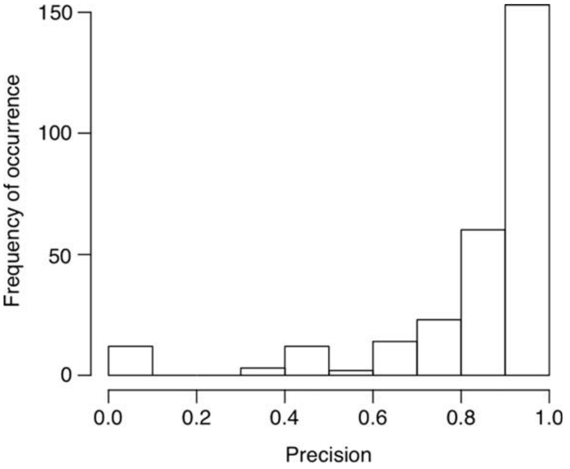

Fig. 4. Actual user queries submitted to PubMed and date limited to 1994 showing the distribution of precision values for highly cited results post-threshold.

SCI categories for oncology and psychiatry. Negative coefficients that were statistically significant included squared JIF term, total number of past publications, both maximum and minimum number of authors' past publications, and the SCI categories for anesthesia and general medicine. JIF and number of authors had the highest positive coefficients. We used these two predictors to develop our filter (see Table 1).

4.2. Sequential result refinement

We ran our hill climber on the 1994 training set. We tested a range of decay parameters and selected 0.8 resulting in a mean precision of 0.886 (95% C.I.: 0.861, 0.911). The thresholds the hill climber identified for articles with five or more citations were JIF (> 7.73), MPACT $(> 9.9 \times 10^{-5})$, and number of authors (> 8).

The 1994 test set contained 56,376 articles from journals with available JIFs. 21,698 articles (38.5%) were cited five or more times. The articles in the 1994 test set had a mean JIF of 3.789 (range: 0.013–52.28; median: 2.375), mean MPACT score 0.012 (range: 0–0.337; median: 0.005), and an average of 3.8 authors (range: 0–64; median: 3). Articles published in journals with JIFs were more likely to be cited than a randomly selected MEDLINE article. When filtered with the identified thresholds we obtained a set of 417 articles, 86.3% of which were cited five or more times, a significant increase in the prevalence of highly cited articles $\chi^2(1, N = 56,766) = 396.85$, $p < 0.01$.

To determine whether our methods developed using 1994 data were stable over time, we applied the same thresholds to the 1999 set. Of 323,806 articles in the 1999 set, 40.5% had five or more citations. The mean JIF was 3.138 (range: 0.03–40.782; median: 1.948), mean MPACT score was 0.013 (range: 0–0.339; median: 0.005), and average number of authors was 4.182 (range: 0–61; median: 4). When the 1999 set was filtered using the identified thresholds, we obtained 2853 articles, 77.1% cited five or more times, $\chi^2(1, N = 326,350) = 1569.35$, $p < 0.01$. Please note that, due to the nature of our data set, we only had five years of citation information for the 1999 set. Our citation data was therefore incomplete, which may have led us to underestimate the percentage of highly cited articles.

4.3. Evaluation using actual PubMed queries

The 426 queries in our random sample had a mean of 3367 results published in 1994 (range: 1–441,742; median: 198). These queries had a mean of 1056 highly cited results published in 1994 (range: 0–108,412; median: 69). The baseline precision for this set of 426 queries was therefore 0.377 (95% C.I.: 0.361, 0.393). After applying the thresholds, out of 279 queries with results (147, 34.5% had no results post-threshold), the mean was 25 results (range: 0–2097; median: 2) (Fig. 4). The majority of queries that had no results post-threshold were queries that originally had fewer than 500 results. There were 12 queries post-threshold that had false positive results returning one or two non-cited articles. After thresholding, the result sets were significantly enriched for highly cited articles with a mean precision of 0.854 (95% C.I.: 0.827, 0.882) compared to 0.377 (95% C.I.: 0.361, 0.393), paired Wilcoxon signed-rank test $(V = 37,817)$, $p < 0.01$. The post-threshold set contained fewer articles; most post-threshold result sets had fewer than 100 articles. The mean decrease in size of the result set post-threshold was 3342 (range: 1–439,600; median: 196). As expected, recall was poor with a mean of 0.027 (95% C.I.: 0.022, 0.032). This recall calculation should be considered an estimate, as the SCI is not comprehensive over all of MEDLINE (we considered highly cited articles 'relevant'). The number of false negatives, highly cited articles published in 1994 missed by this method, had a mean of 1033 results (range: 0–106,600; median: 68). One query, "bar or," missed 106,600 articles.

We were concerned that 34.5% of queries did not return any results. This could have been an artifact of restricting to 1994 data. We therefore resubmitted the 426 queries to PubMed with a date restriction of 1980–2006 and evaluated whether the method would return any results. Of the 426 queries, only 31 (7.3%) had no results after applying the thresholds.

5. Discussion

We developed a method to filter large result sets to select articles that are likely to be highly cited in the future. We accomplished this by determining thresholds for three simple predictors available at the time of listing in MEDLINE: JIF, MPACT, and number of authors. When these thresholds were applied to PubMed query results, the resulting subsets were significantly enriched with highly cited articles. Thresholds developed using 1994 data were still effective for 1999 data. For PubMed queries with more than 1000 results, this technique reduced the number of articles to be reviewed, and could be used to rank newly published articles for which there is no citation information.

Our method has several important limitations. First, it still requires the JIF, which is derived from citation information. Inclusion in citation databases (and an official JIF) is desirable for a journal. This means that articles submitted to biomedical journals with JIFs are more likely to attract citations than articles submitted to other journals. For example, the prevalence of articles with five or more citations was much higher in our 1994 data set than in the general article population (38.5% versus 2%) [14]. Therefore our results may not generalize to articles in journals not covered by the SCI. The mapping of SCI to MEDLINE was imperfect covering 33% of articles. The SCI also does not cover journals outside of the US or Europe, and in languages other than English. However, we believe that this is not a major impediment for the majority of searchers who desire important articles. It is likely that high quality articles can be found in highly cited journals with large JIFs.

Second, our method did not locate all articles with five or more citations. In fact this method missed on average 1000 highly cited articles from a random set of queries, locating 2–3% of all highly cited articles. Our variables were filtered in a sequential fashion to rapidly scale down on the number of results. Since we sacrificed recall for precision, these filters would aid users that require just a few highly cited articles, such as students [4]. Users that require comprehensive results should use other techniques. For example, a system could present all results but highlight articles that passed our filter as likely to be highly cited in the future.

Third, MeSH term assignment is associated with a short delay from publication in a journal. Articles with "in-process" MEDLINE records are listed and retrievable by PubMed, but only have terms assigned by the publisher. MeSH term assignment by trained indexers can take up to eight weeks. We do not know how much this phenomenon affects newly listed articles as it was not covered by our experimental design. Since in-progress records are, at any given time, a small percentage of MEDLINE records, it is unlikely that result sets filtered using our technique would be greatly affected. However, this time delay and the implied changes in the result sets once records are indexed is an additional limitation.

Our definition of MPACT used the calendar year as a time cutoff. Definitive MPACT scores are therefore not available for the ongoing year. In the worst-case scenario, articles published in January would have to wait almost an entire year for their MPACT scores to be obtainable. This would limit our goal of using information available at the time of listing in MEDLINE. There are several potential workarounds. For example, MPACT scores for the current year could simply be recomputed every time MEDLINE is updated, using the count of all published articles to that point in time or other time-series approaches.

Finally, this method is a classifier and does not rank articles. All articles that pass the filter have the same rank, because we simply classified articles into two broad categories: articles with five or more citations in 10 years and those without. To develop a ranking process for search results, we will work on methods that can categorize articles into multiple classes. For example, techniques such as decision trees may improve prediction granularity.

An alternative article ranking strategy relies on implicit feedback, such as clickthrough data [39]. A clickthrough is a user action on a link to view the article's abstract or full text. Clickthroughs are assumed to correlate with user interest in the article. A MEDLINE interface such as PubMed could track the number of clickthroughs. As these data accumulate, they can be used to rank articles. However, such a strategy would require time for clickthroughs to accumulate. Our method can bridge the gap from publication until sufficient clickthroughs accumulate to allow meaningful ranking. Of course, if a certain article is identified as likely to be important and is placed high in the result list, it will likely garner clickthroughs faster than the same article lower in the result set.

A scoring system with known variables may be manipulated to maximize an author's chance to be cited. Just like journals with a high JIF are preferred by authors, if our predictors become widely accepted authors may try to influence the system by manipulating these variables to make their articles pass our filter (i.e., creating a self-fulfilling prophecy). Fortunately, this will be difficult. The only input to our filter that can be purposefully manipulated is the number of authors. Both the JIF score and the assignment of MeSH terms are controlled by independent third parties. The JIF is time limited and based on the number of citations to articles in a journal. MeSH terms are assigned by trained human indexers employed by the NLM. We doubt that authors will pad the author list for their articles since it violates the authorship rules of the International Committee of Medical Journal Editors (ICMJE) and would dilute their own contribution [40].

In future work, we will identify additional predictors that are less dependent on JIF, further reducing our dependence on a citation database. We will develop ways to generalize our work by leveraging the MeSH hierarchy. We will also explore converting MPACT to stemmed text-words or the output of an automated classifier: eliminating the time lag for MeSH term assignment. Time-series methods or the use of a sliding time window may also improve MPACT calculation. Although this project focused on MEDLINE, our approach may be applicable to other fields once we are able to generalize the MPACT score as the SCI tracks citations in fields other than biomedicine.

Our ultimate goal remains a search interface that presents important work first. A highly cited article does not necessarily mean that it is high quality, relevant, or important. Thus, expert-selected gold standards remain extremely valuable for this line of research. Since the true test of any information retrieval system is whether it can answer the information needs of its users, a user-based study comparing various search methodologies will determine whether other filters truly benefit users.

6. Conclusions

Science continuously produces new information in vast quantities. Citation analysis can help identify highly cited articles, and can be applied to rank articles within large result sets. However, citation databases are hard to create, maintain, and suffer from citation lag. We developed a method to select a subset of articles likely to be highly cited five or more times using three predictors: JIF, MPACT, and the number of authors. Further research is needed to find other predictors to identify a subgroup of newly published articles that are likely to influence the field.

References

[1] US National Library of Medicine. NLM systems: data, news and update information. Retrieved February 24, 2007. Available from: http://www.nlm.nih.gov/bsd/revup/revup_pub.html; February 2007.

[2] Herskovic JR, Tanaka LY, Hersh WR, Bernstam EV. A day in the life of PubMed: analysis of a typical day's query log. J Am Med Inform Assoc 2007;14(2):212–20.

[3] Silverstein C, Marais H, Henzinger M, Moricz M. Analysis of a very large web search engine query log. ACM SIGIR Forum 1999;33(1):6–12.

[4] Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? JAMA 1998;280(15):1347–52.

[5] Bernstam EV, Herskovic JR, Aphinyanaphongs Y, Aliferis CF, Iyengar MS, Hersh WR. Using citation data to improve retrieval from MEDLINE. J Am Med Inform Assoc 2006;13(1):96–105.

[6] US National Library of Medicine. Pubmed clinical queries. Retrieved April 29, 2008. Available from: http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.shtml; 2008.

[7] UpToDate: welcome. Retrieved July 10, 2007. Available from: http://www.uptodateinc.com/; 2007.

[8] Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR. Optimal search strategies for retrieving scientifically strong studies of treatment from MEDLINE: analytical survey. BMJ 2005;330(7501):1179.

[9] 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(2):207–16.

[10] Aphinyanaphongs Y, Statnikov A, Aliferis CF. A comparison of citation metrics to machine learning filters for the identification of high quality MEDLINE documents. J Am Med Inform Assoc 2006;13(4):446–55.

[11] Garfield E. Can citation indexing be automated? In: Stevens ME, Giuliano VE, Heilprin LB, editors. Statistical association methods for mechanized documentation. Symposium proceedings, Washington 1964. National Bureau of Standards Miscellaneous Publication 269; 1965. p. 189–92.

[12] Garfield E. Citation analysis as a tool in journal evaluation. Science 1972;178(4060):471–9.

[13] Kleinberg J. Authoritative sources in a hyperlinked environment. In: Proceedings ninth ACM-SIAM symposium on discrete algorithms; 1998.

[14] de Solla Price DJ. Networks of scientific papers. Science 1965;149(3683):510–5.

[15] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the seventh international conference on world wide web, vol. 7; 1998. p. 107–17.

[16] Burrell QL. Predicting future citation behavior. J Am Soc Inform Sci Technol 2003;54(5):372–8.

[17] Garfield E. The Thomson scientific impact factor. Retrieved December 5, 2007. Available from: http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor; June 1994.

[18] Garfield E. Which medical journals have the greatest impact? Ann Intern Med 1986;105(2):313–20.

[19] Garfield E. Panel on "Evaluative measures for resource quality: Beyond the impact factor". In: Medical library association meeting. Medical Library Association; 2007.

[20] Seglen PO. Why the impact factor of journals should not be used for evaluating research. BMJ 1997;314(7079):498–502.

[21] Van Dalen HP, Henkens K. What makes a scientific article influential? The case of demographers. Scientometrics 2001;50(3):455–82.

[22] Newman MEJ. The structure of scientific collaboration networks. Proc Natl Acad Sci USA 2001;98(2):404–9.

[23] Leimu R, Koricheva J. Does scientific collaboration increase the impact of ecological articles? BioScience 2005;55(5):438–43.

[24] Kuhn TS. The structure of scientific revolutions. 3rd ed. Chicago, IL: The University of Chicago Press; 1996.

[25] Chen C. Citespace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inform Sci Technol 2006;57(3):359–77.

[26] Kleinberg J. Bursty and hierarchical structure in streams. In: Proceedings eighth ACM SIGKDD international conference on knowledge discovery and data mining; 2002.

[27] Ljosland M. Evaluation of web search engines and the search for better ranking algorithms. Retrieved August 3, 2007. Available from: http://citeseer.ist.psu.edu/ljosland99evaluation.html; 1999.

[28] Hawking D, Craswell N, Bailey P, Griffiths K. Measuring search engine quality. Inform Retriev 2001;4(1):33–59.

[29] Python Software Foundation. Python programming language—official website. Retrieved July 8, 2007. Available from: http://www.python.org; 2007.

[30] BioPython. BioPython. Retrieved July 8, 2007. Available from: http://www.biopython.org; 2007.

[31] US National Library of Medicine. MEDLINE/PubMed data element (field) descriptions. Retrieved March 21, 2007. Available from: http://www.nlm.nih.gov/bsd/mms/medlineelements.html; 2006.

[32] PostgreSQL Global Development Group. PostgreSQL. Retrieved July 8, 2007. Available from: http://www.postgresql.org; 2007.

[33] Cameron AC, Trivedi PK. Regression analysis of count data. Cambridge: Cambridge University Press; 1998.

[34] US National Library of Medicine. Medical subject headings—files available to download. Retrieved October 9, 2007. Available from: http://www.nlm.nih.gov/mesh/filelist.html; October 2007.

[35] Funk ME, Reid CA, McGoogan LS. Indexing consistency in MEDLINE. Bull Med Librarian Assoc 1983;71(2):176–83.

[36] Russell SJ, Norvig P. Artificial intelligence: a modern approach. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 2002.

[37] US National Library of Medicine. Index of ftp://ftp.ncbi.nlm.nih.gov/toolbox/pubmed/query-logs/. Retrieved October 17, 2005. Available from: ftp://ftp.ncbi.nlm.nih.gov/toolbox/pubmed/query-logs/; October 2005.

[38] The R project for statistical computing. Retrieved July 8, 2007. Available from: http://www.r-project.org; 2007.

[39] Joachims T. Optimizing search engines using clickthrough data. In: KDD'02: proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY, USA: ACM; 2002. p. 133–42.

[40] International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication. Retrieved January 14, 2009. Available from: http://www.icmje.org; October 2008.