Updating a bibliography using the RELATED ARTICLES function within PubMed

Xiaole Liu and Russ B. Altman, MD, PhD,
Stanford Medical Informatics, Stanford University Medical Center,
MSOB X-215, Stanford, CA 94305-5479, USA
{xliu, altman}@smi.stanford.edu

Abstract

Comprehensive bibliographies are useful for conducting reviews of the literature, and for assessing the progress within a field. These bibliographies may be broad and inclusive, or focused and precise in their inclusion criteria. In either case, the task of maintaining a complete bibliography within a particular area of research is made difficult by the diversity, complexity and huge volume of newly published literature. In an effort to effectively and automatically retrieve relevant literature, different search strategies and indexing tools have been developed, including the RELATED ARTICLES function provided with the PubMed[1] system. In this paper, we report a program for incremental updates of a bibliography using the PubMed RELATED ARTICLES function. Given a highly specialized starting bibliography of experimental measurements of the structure of the 30S bacterial ribosomal subunit, the system was applied to find additional relevant references. For this particular task, the system has a recall of 75%, a strict precision of 32% and a partial precision of 42%. Our results are notable because although the RELATED ARTICLES function is purely statistical, it is nonetheless able to select a very narrowly defined set of articles from the literature. We discuss the tradeoffs between having a user to evaluate many articles of possible interest in a single session, versus asking a user to evaluate a small set of articles on a periodic basis.

Introduction

The maintenance of a comprehensive bibliography relevant to a specialty area is a recurring challenge within medical research. Such bibliographies play a role in periodic reviews, assessments of overall progress and for historical analysis. In addition, such bibliographies are also used at the beginning of a project in order to assess the data available to answer a particular question. The task of gathering the relevant literature has been simplified greatly with the general availability of a computer-based literature-searching engine for the medical literature, MEDLINE[2]. Nonetheless, constructing a comprehensive bibliography often requires a very broad search, followed by a meticulous review of hundreds or thousands of references in order to identify the most relevant reports.

The strategies that are used for assembling a comprehensive bibliography naturally make use of the existing capabilities of literature retrieval systems. There are a number of useful heuristics for updating a bibliography, and these include:

(1) Updates using the authors' names. Most authors publish in the same area repeatedly, and so this heuristic allows the progress of particular investigator's to be tracked. Of course, investigators also change interests or evolve their focus, and so this is not always a reliable method.

(2) Updates using citation profiles. The Science Citation Index[3] maintains a record of all the articles that refer to a given article after its publication. Citation profiles therefore can be used to find articles relevant to a bibliography by chasing these forward links[4]. However, since all relevant work in an article has to be cited, including basic background material and incidental experimental and scientific details, the precision of the references with respect to particular inclusion criteria is not always high.

(3) Updates using keyword appearance. The use of particular "pathognomonic" words in a literature is useful for identifying reports based on the occurrence of specific words, or MeSH concepts[5]. However, indexing by word appearance may produce a huge list of references with low precision. In addition, the reliability of consistent word usage within a field is not high[6].

(4) Updates using a controlled vocabulary[7]. Using natural language processing techniques or manual indexing, a free-text document can be mapped into symbolic representation based on a predefined controlled vocabulary, such as UMLS metathesaurus[8, 9, 10, 11]. The resulting labels can be used to find relevant literature[12], but the assignment of terms may be biased or inconsistent[13].

(5) Updates using word frequency and weight. Recent work has shown that mapping text into concepts has little or no advantage over statistical word-based systems which compute the relevance between two articles by counting the frequency of

1091—8280/98/$5.00 © 1998 AMIA, Inc.

shared text words in the two articles[14, 15, 16, 17]. There are also techniques for weighting text words based on their ability to convey meaning[18]. The PubMed RELATED ARTICLES function is a word-based indexing tools that consider both frequency and weight of text words.

In principle, the power of all these methods is increased when a larger set of articles is available as a starting point, instead of a single article, or a set of target keywords. Sets of closely related articles provide a larger training set for retrieval methods, and allow the noise in all these methods to be averaged out. In this paper, we tested our ability to use the RELATED ARTICLES feature of PubMed to incrementally add relevant articles to a highly specific bibliography. PubMed is the web interface to the MEDLINE database which contains over 8 million references from about 3,800 biomedical journals. Each reference has a unique identification number in both MEDLINE (UI) and PubMed (PMID). PubMed offers a RELATED ARTICLES function[19] which links each article to a list of references that score highly using a relevance ranking. The algorithm represents a document by all the words in the title, abstract and MeSH terms, filtering out a set of common uninformative words. Within PubMed, each non-noise word has a global weight determined by the number of references in the database that contain the word and how strongly it is associated with the assigned MeSH headings. Each word also has a document-dependent local weight which is the number of times it occurs in the title, abstract and MeSH terms of the document. When two articles are compared, the product of (global wt * local wt1 * local wt2) of all shared words are summed to produce a total weight. This sum is divided by the product of the lengths of the two documents to adjust for document length. The RELATED ARTICLES list reports the high scoring articles, in descending score order. This list is precomputed for each reference in PubMed, so it can be retrieved quickly when the RELATED ARTICLES button is pushed. This algorithm is efficient in both time and memory with a precision of 40% in test studies[20].

Our comprehensive bibliography is in the area of the structure of the 30S ribosomal subunit of bacteria. The ribosome is where the DNA code is translated into protein, and the structure of the ribosome has been studied by many groups for over thirty years. We are building a knowledge base of the published literature about the 30S subunit which requires that we identify all articles with structural information dealing with this subunit. The task is difficult because we are specifically interested in structural measurements, as opposed to other functional information. In addition, the bacterial ribosome has another subunit, the 50S subunit, which is not currently our focus. Other organisms, such as yeast and human also have ribosomes that are the subject of intense study, and these ribosomes are not currently of interest to us. We created our core bibliography by finding five review articles published between 1984 and 1996 and taking the union of their bibliographies. The resulting core bibliography contained 206 journal articles of potential interest. We reviewed each of these and classified them into five categories: (1) directly relevant to the structure (87 articles), (2) indirectly relevant to structure (35 articles), (3) relevant to the function of ribosomes (28 articles), (4) general molecular biology references (43 articles) and (5) distantly related to the subject of interest (13 articles). Although this was a relatively complete bibliography, we needed a way both to find old articles that were missed in the union of bibliographies, and to find new articles as they appear in the literature. In order to give our algorithms "partial credit" for references that contain information on the closely related 50S subunit in bacteria, the yeast or the human ribosome, we used the same distribution into 5 categories for classifying these articles, but gave them relevance codes of 100 + category (for 50S), 200 + category (for the eukaryotic ribosomal data). Using the RELATED ARTICLES function, we developed an incremental bibliography updating system. Our system is incremental in the sense that it is run periodically and suggests a specified number of new articles each time it is run, instead of proposing hundreds of new articles.

Methods

The basic strategy of the system is as follows. From a starting bibliography (in this case, the set of 87 articles classified as directly relevant), the system uses the PubMed unique ID and a URL call to retrieve a list of RELATED ARTICLES for each article in the bibliography. These articles are then scored, based on their frequency of occurrence on the list of RELATED ARTICLES for the set of starting references. We have created two scoring strategies. In the *simple count* approach, we simply count the number of times an article is linked to one of the core references within the bibliography, without reference to its rank within the list returned by PubMed. In the *linear weight* approach, if a core bibliography reference finds a list of *s* RELATED ARTICLES (ranked by PubMed), we increment the score of the first article *by s* and the score of the r^{th} article by *(s-r)*. With both these scoring systems, if an article is linked to many core articles, it will have a higher score. Articles with highest scores are reported to the user. For each newly found article, the system reports its PMID, UI, score, title and its web link to the PubMed query which contains the abstract. The user is asked to evaluate the proposed articles (a parameter can be set to indicate how many articles the user is to evaluate in each iteration) and indicate which belong in the bibliography. The PMID of articles that are added are maintained and the system is ready for the second layer of updating. After each iteration, the newly discovered articles are added to the core list and the bibliography is updated. For each reference reported to the user or present in the starting bibliography, its PMID is stored so it will not be presented by the system again.

The automated biomedical bibliography updating system is written in Perl to run on a personal computer. Its performance was tested in four ways. First, we performed an exhaustive analysis of the articles linked to the core bibliography of 87 articles. For each of these articles we scored the top 40 RELATED ARTICLES reported by PubMed and created a database with 1090 articles (the union--with duplicates removed--of the 87 lists of 40 articles each), reported in random order without the scoring information. We then manually went through the titles and abstracts of these 1090 articles and classified them into the five categories described above. These ranks were used as the gold standard to compare to the score assigned by the system. Articles of rank 1 represented *strict* relevance to the structural model of small 30S subunit of prokaryotic ribosome. Articles of rank 101 (about the large subunit of prokaryotic ribosome) and 201 (about the eukaryotic ribosome) represented *partial* relevance to the structural model of ribosome, since they are conceptually very closely related and differ only in the species being studied. All other ranks (2-5, 102-105, 202-205) denote less relevant references.

Second, we performed cross validation. From the starting bibliography, one article of rank 1 was taken out at a time to see whether the other 86 could retrieve it from PubMed RELATED ARTICLES (using the top 15 reported articles) based on its appearance in the total list of RELATED ARTICLES for the remaining 86. In addition, we measured how many articles within the larger ribosome bibliography (206 total) could be found by these highly relevant articles.

Third, we tested the efficacy of an incremental, layered updating system, and its sensitivity to the key parameters. The parameters tested were: the number of RELATED ARTICLES retrieved for each core reference, and the number of scored articles reported to the user for evaluation. We evaluated the partial precision of each strategy. The system was run for 10 iterations. In each iteration, the rank 1 articles were selected by the user, and added to the core bibliography for the next round of search. We compared strategies of reporting 15, 20, 40 and 80 articles to a user for manual evaluation of relevance. We also compared the performance of our two scoring functions.

Finally, we tested the performance of the method as a function of the size of the initial corpus. We tested initial bibliographies of size 10, 20, 40 and 80, randomly selected from the core collection of 87 articles, and measured precision.

For these experiments, we computed *strict precision* as the percent of retrieved articles that are given a rank of 1 in the gold standard classification of 1090 articles. *Partial precision* is the percent of retrieved articles that are given a rank of 1, 101 or 201 in the classification. Recall is the percentage of all rank 1 articles retrieved by the system divided by all rank 1 articles in the document collection (starting bibliography)[21].

Results

Among the total of 1090 articles generated from a list of 40 RELATED ARTICLES for each of the core 87 articles, 146 additional articles were rank 1 and 190 were rank 101 or 201, and so we can assign an overall strict precision of 13.4% and partial precision of 30.8% to the straightforward application of the RELATED ARTICLES functionality. We used the ranks of these 1090 articles as a gold standard for comparison with our scoring methods, discussed below.

In the cross validation experiment, the missing article was retrieved 65 out of 87 times when each of the remaining references was used to find 15 RELATED ARTICLES. This represents a recall performance of 74.7%. The 87 core articles find only 39 of the remaining 206 - 87 = 119 non-core articles in the original ribosome bibliography, which represented the union of the bibliographies of five review articles.

In comparing the simple count scoring method and the linear weight approach, we found that the linear weight approach consistently outperformed the simple count method with a difference in precision ranging between 10% to 40%. The average strict precision for the linear weight approach is 32% (42% partial precision), while the simple count approach has an average strict precision of 27% (38% partial precision).

Table I summarizes the ability of the system to rank relevant articles highly. If twenty articles are reported in each round of incremental updating, then half of these articles are strictly relevant and 60% are partially relevant. As more articles are reported, the user is asked to do more work evaluating relevance, but the yield decreases. With one hundred articles, 32% are strictly relevant. Thus, a user can retrieve 32 strictly relevant articles by either evaluating three groups of twenty articles in three sessions, or by evaluating one group of 100 articles in a single session. The performance of the method is better with multiple sessions because the bibliography is updated, and so performance in subsequent iterations remains high.

Table I: Ability of the scoring functions to rank strictly relevant articles. The first column reports the number of articles reported to the user. The second column shows the percentage of retrieved articles that are most relevant to the bibliography. The third column shows the percentage of retrieved articles that are reasonably relevant (categories 1, 101 and 201 from the text) to the bibliography.

# of articles reported	% strictly relevant	% partially relevant
20	50%	60%
40	43%	53%
100	32%	41%
278	30%	46%

Table II summarizes the performance of our method relative to the size of the starting bibliography. We took random subsets of the core bibliography of 87 articles of size 10, 20, 40 and 80 and computed the strict and partial precision for the first cycle of incremental updating. Thus, when only 10 articles are available, 30% of the reported references are scored at rank 1. Up to 80 articles, there is still a trend of increasing relevance, indicating that a large starting bibliography is most useful for acquiring new articles. We know, based on our gold standard evaluation of 1090 articles, that there are at least 146 + 87 = 233 strictly relevant articles.

Table II: Performance of the system with different number of references in the starting bibliography. The first column shows the size of the starting bibliography, used for search. The second and third columns are as described in Table 1.

Size of core bibliography	% strictly relevant	% partially relevant
10	30%	35%
20	40%	40%
40	45%	50%
80	50%	55%

Discussion

Our gold standard of 1090 articles is not perfect. It is generated by using the RELATED ARTICLES function, and therefore is biased by the characteristics of that algorithm. We used a large fraction of the reported links in order to maximize the chance of capturing useful articles. Our cross-validation evaluation of recall indicates that only 75% of the strictly relevant articles can be recalled with the top 15 RELATED ARTICLES. The articles that were missed in the cross-validation often appeared in PubMed with no abstract, and tended to be the older articles.

Our experiments are a particularly difficult test of the utility and limitations of using the PubMed RELATED ARTICLES link for expansion of a bibliography. Our ribosomal bibliography is highly specialized, and contains references selected because they contain structural information about one subunit of a particular organelle within a particular organism. Articles dealing with identical experimental approaches applied to related subunits, organelles or organisms are not considered "strictly relevant." There is substantial heterogeneity in the terminology within the field, and so this subunit can be termed the *16S ribosomal subunit*, the *30S ribosomal subunit*, the *bacterial* or *prokaryotic small subunit*, or even the *small part of the 70S subunit.* In addition, the papers of interest provide structural information (as opposed to functional or genetic) but the names of the experiments do not provide strong clues as to the nature of the data. As such, we believe our particular bibliographic updating task is a very difficult one. Perhaps the most compelling evidence that we have a difficult bibliography to reproduce, is our finding that the core group of 87 highly selected articles contained 22 articles that had no "related article" link to any of the other article in the group. There are MeSH headings for some of the topics associated with our bibliography, but these are not consistently used in our bibliography, and are often at a coarse grain compared to the inclusion criteria of our bibliography.

Nevertheless, we are able to draw some conclusions about the task of updating a bibliography. First, an incremental approach may be more acceptable both in terms of performance and acceptability. With respect to performance, we clearly showed that by proposing a small number of highly ranked new articles, and allowing a user to evaluate them and add a subset (often as high as 50% of the proposed articles), we can make incremental improvements to the bibliography which allow it to continue to propose useful articles at least through the ten incremental steps we tested. At the same time, it is clearly much more palatable for a user to be presented with twenty articles every few days to evaluate, rather than 1090 all at once. Especially if the user is curating a collection of articles over time, and does not face a deadline for updating the bibliography, this gradual method is easy to implement with E-mail and allows constant surveillance of the literature.

We also can draw some conclusions about the utility of the RELATED ARTICLES functionality in PubMed. First, there is clearly a correlation between the strict relevance of an article and its rank in the RELATED ARTICLES list. Our linear weight scoring method (which rewarded articles at the top of the RELATED ARTICLES list) clearly outperformed the simple count approach. In addition, we found that by expanding the number of articles taken from that list, we degraded precision significantly (as shown in Table I). The low scoring articles in the RELATED ARTICLES list often are general references dealing with the same field, or are unrelated articles that by chance share some unusual words. The top twenty to forty articles contain the most useful and relevant set of links.

Our incremental approach currently depends solely on the RELATED ARTICLES link. It is quite probable that a combined method that uses forward and backward references and the names of authors would yield a more complete system with improved recall, and we are considering strategies for a hybrid approach. The method as we have implemented it now takes advantage of the free PubMed database, and combines information from keywords and MeSH, through its use of the *RELATED ARTICLES* link. Over time, the incremental approach should find most articles of interest and decline in performance. Our study is limited in that we did not run the incremental approach enough times to evaluate how many increments are required to find the complete set of 233 articles known to be strictly relevant. Initially, our core group of 87 articles reports twenty possibly relevant articles with a success rate of about 50%. This rate decreases over time, and we are currently evaluating the number of cycles required to reproduce all known relevant articles.

Acknowledgements

RBA is supported by NIH LM-05652, LM-06422, NSF DBI-9600637 and a grant from IBM. Aldo Calpini provided useful Perl code.

References

1 PubMed: http://www.ncbi.nlm.nih.gov/PubMed/

2 The NLM PubMed Project. http://www.ncbi.nlm.nih.gov PubMed/overview.html#Medline

3 SciSearch® at LANL: http://scisearch.lanl.gov/

4 Shalini R. 'Citation profiles' to improve relevance in a two-stage retrieval system: a proposal. Information Processing & management 1993 July-Aug;29(4):463-70.

5 PubMed Help: http://www.ncbi.nlm.nih.gov/PubMed/pubmedhelp.html

6 Efthimiadis EN, Afifi M. Population groups: indexing, coverage, and retrieval effectiveness of ethnically related health care issues in health sciences databases. Bull Med Libr Assoc 1996 Jul;84(3):386-96.

7 Rada R, Blum B, Calhoun E, Mili H, Orthner H, Singer S. A vocabulary for medical informatics. Comput Biomed Res 1987 Jun;20(3):244-63.

8 Wagner MM. An automatic indexing method for medical documents. Proc Annu Symp Comput Appl Med Care. 1991:1011-7.

9 Chute CG, Yang Y, Evans DA. Latent Semantic Indexing of medical diagnoses using UMLS semantic structures. Proc Annu Symp Comput Appl Med Care. 1991:185-9.

10 Hersh WR, Greenes RA. SAPHIRE--an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. Comput Biomed Res 1990 Oct 23(5):410-25.

11 Detmer WM, Barnett GO, Hersh WR. MedWeaver: integrating decision support, literature searching, and Web exploration using the UMLS Metathesaurus. Proc AMIA Annu Fall Symp. 1997:490-4.

12 Chute CG, Yang Y. An evaluation of concept based latent semantic indexing for clinical information retrieval. Proc Annu Symp Comput Appl Med Care. 1992:639-43.

13 Hersh WR. Evaluation of Meta-1 for a concept-based approach to the automated indexing and retrieval of bibliographic and full-text databases. Med Decis Making 1991 Oct;11(4 Suppl): S120-4.

14 Hersh WR, Hickam DH, Leone TJ. Words, concepts, or both: optimal indexing units for automated information retrieval. Proc Annu Symp Comput Appl Med Care. 1992:644-8.

15 Hersh WR, Hickam DH. A comparison of retrieval effectiveness for three methods of indexing medical literature. Am J Med Sci 1992 May;303(5):292-300.

16 Yang Y, Chute CG. Words or concepts: the features of indexing units and their optimal use in information retrieval. Proc Annu Symp Comput Appl Med Care. 1993:685-9.

17 Hersh WR, Hickam DH. A comparison of two methods for indexing and retrieval from a full-text medical database. Med Decis Making 1993 Jul;13(3):220-6.

18 BH Weinberg, JA Cunningham. Word frequency data in full text database searching. National Online Meeting Proceedings 1984 Apr 10-12: New York, NY, USA; 1984. p. 425-32.

19 Computation of RELATED ARTICLES. URL: http://www.ncbi.nlm.nih.gov/PubMed/computation.html

20 Wilbur WJ, Yang Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Comput Biol Med 1996 May;26(3):209-22.

21 Purcell GP. Contextual document models for searching the clinical literature [dissertation]. Stanford (CA): Stanford Univ. 1996.