BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 7 2009, pages 974–976
doi:10.1093/bioinformatics/btn033

Data and text mining

MiSearch adaptive pubMed search tool

David J. States[1,2,*], Alex S. Ade[1], Zachary C. Wright[1], Aaron V. Bookvich[1] and Brian D. Athey[1,3]

[1]National Center for Integrative Biomedical Informatics, [2]Department of Human Genetics and [3]Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109, USA

Received on September 7, 2007; revised and accepted on January 22, 2008

Advance Access publication March 6, 2008

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: MiSearch is an adaptive biomedical literature search tool that ranks citations based on a statistical model for the likelihood that a user will choose to view them. Citation selections are automatically acquired during browsing and used to dynamically update a likelihood model that includes authorship, journal and PubMed indexing information. The user can optionally elect to include or exclude specific features and vary the importance of timeliness in the ranking.

Availability: http://misearch.ncibi.org

Contact: dstates@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

With the rapidly increasing volume of publications in the biomedical literature, finding relevant work is an ever more difficult challenge. General solutions to the literature search problem are difficult because biomedical science is very diverse; the articles most relevant to one reader may not be relevant to another. Relevance feedback is a well-established technique to improve performance in information retrieval (Rocchio, 1971; Salton, 1971; Salton and Buckley, 1990). Feedback may be acquired explicitly by asking users to rate retrieval results. However, many users find this task burdensome. Even for widely deployed search engines such as Excite, where relevance feedback is available and effective, it is rarely used (Spink *et al*., 2000). An alternative is to acquire feedback implicitly by observing user behavior (Kelly and Teevan, 2003).

MiSearch is an adaptive literature search tool using implicit relevance feedback that helps users rapidly find PubMed citations relevant to their specific interests. MiSearch automatically saves information on citations a reader has viewed during search and browsing, and uses this information to build a statistical profile describing the readers' choices. This profile is used to rank the results of future searches, placing those articles that this reader is most likely to view at the top of the list. In effect, MiSearch is using query expansion with probabilistic weighting of terms derived from the implicitly defined relevant document set. Using this implicit feedback approach is effective and improves the relevance ranking of bibliographic search results.

The NCBI Entrez search tool is widely used and alternative interfaces have been developed allowing users to manually vary the weight of different features in determining relevance (Muin *et al*., 2005) and to reformulate and refine Boolean queries (Bernstam, 2001; Ding *et al*., 2006), but unlike MiSearch, these tools do not adapt to user behavior.

2 METHODS

2.1 Ranking algorithm

MiSearch records the users search history and the history of documents selected for viewing using an HTTP redirect mechanism. Four domains are considered: authors (Au), journal (Jl), MeSH terms (Me) and substance names (Sn) indexed by NLM (Nelson *et al*., 2004). Each domain is described using a statistical profile of term use. The frequency $f_u(t)$ of term t occurring in citations that user, u, has selected for viewing is defined as

$$f_u(t) = \frac{(N_u(t) + f_P(t))}{N_u + 1}$$

where $N_u(t)$ is the count of citations indexed with term t that were viewed by the user,N_u is the total number of citations viewed by the user and $f_P(t)$ is the absolute frequency with which papers indexed with term t occur in the entire PubMed database. The pseudo counts smooth behavior when the profile has few citations and avoid division by zero if a specific term does not occur in the citations selected by a user. If no feedback is available for the user ($N_u = 0$), then $f_u(t)$ is $f_P(t)$. When the user has viewed many articles, $f_u(t)$ asymptotically approaches $N_u(t/N_u)$.

MiSearch uses the PubMed eUtils interface to query the PubMed database and ranks citations based on a log likelihood score, S,

$$S = \sum_{D=\text{Au, Jl, Me, Sn}} S_D + \alpha(T - T_0)$$

where S_D are log likelihood scores for each domain and $\alpha(T - T_0)$ is term weighting the timeliness of an article. T is the date of publication for a citation, T_0 is a reference date (January 1, 2000) and α is an adjustable factor that allows the user to vary the weight given to timeliness in ranking citations.

The score S_D for domain D is calculated for each citation as a log likelihood ratio that the term t associated with this citation occur in citations viewed by the user, f_u, relative to their frequency in all of PubMed, f_P

$$S_D = \sum_{t \in \text{Term}} \log\left(\frac{f_u(t)}{f_P(t)}\right)$$

*To whom correspondence should be addressed.

A positive score indicates that a citation is more likely to be viewed by the user, and a negative score indicates that a citation is less likely to be selected for viewing.

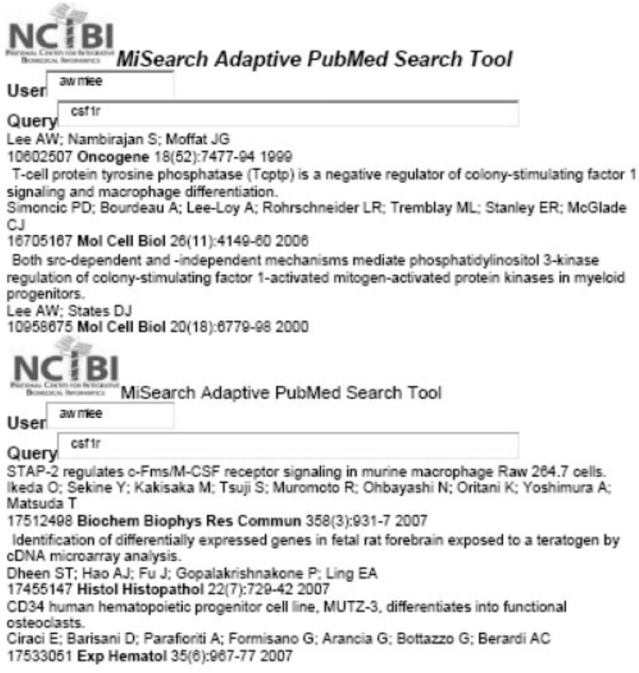
2.2 Implementation

MiSearch is implemented in two components, a PHP script running on an Apache web server that generates forms, dispatches search requests to NCBI Entrez and communicates with a local citation datbase, and a relational database server that stores both the PubMed corpus and user search histories. Ranking is implemented as an SQL stored procedure. Users can label profiles with a string and can define several different profiles for different search tasks.

3 RESULTS

Figure 1 illustrates the effect of adaptive re-ranking of citation searches based on a profile created from the publication of one author (AWL). Dr Lee's research focuses on signal transduction downstream of the CSF-1 receptor (CSF1R). Using a profile based on viewing Dr Lee's publications (top panel), MiSearch ranks two of Dr Lee's publications and a third recent and highly relevant publication at the highest relevance in a PubMed search for 'CSF1R'. In contrast, without adaptive ranking (lower panel), the publications are ranked in reverse chronological order with only one moderately relevant publication highly ranked and that citation is in a journal that Dr Lee does not frequently read.

Because MiSearch ranks citations using a statistical profile, the user does not need to explicitly specify the ranking criteria. MiSearch thus complements Boolean search strategies. In Boolean searches, a relevant article may be missed if the user specifies an overly restrictive Boolean filter and the citation uses a synonym for a term not specified in the search query. Using MiSearch, a broader Boolean query can be performed. The MiSearch relevance ranking places the citations most likely to be of interest to this user at the top of the list and avoids the need to view large numbers of citations. Further, a reader may not be aware that all the citations they are viewing contain a common term such as reference to a chemical substance. The MiSearch statistical profile will automatically capture this information and rank other citations, mentioning this term more highly.



MiSearch Adaptive PubMed Search Tool

User awmee

Query csf1r

Lee AW; Nambirajan S; Moffat JG
10602507 **Oncogene** 18(52):7477-94 1999
T-cell protein tyrosine phosphatase (Tcptp) is a negative regulator of colony-stimulating factor 1 signaling and macrophage differentiation.
Simoncic PD; Bourdeau A; Lee-Loy A; Rohrschneider LR; Tremblay ML; Stanley ER; McGlade CJ
16705167 **Mol Cell Biol** 26(11):4149-60 2006
Both src-dependent and -independent mechanisms mediate phosphatidylinositol 3-kinase regulation of colony-stimulating factor 1-activated mitogen-activated protein kinases in myeloid progenitors.
Lee AW; States DJ
10958675 **Mol Cell Biol** 20(18):6779-98 2000

MiSearch Adaptive PubMed Search Tool

User awmee

Query csf1r

STAP-2 regulates c-Fms/M-CSF receptor signaling in murine macrophage Raw 264.7 cells.
Ikeda O; Sekine Y; Kakisaka M; Tsuji S; Muromoto R; Ohbayashi N; Oritani K; Yoshimura A; Matsuda T
17512498 **Biochem Biophys Res Commun** 358(3):931-7 2007
Identification of differentially expressed genes in fetal rat forebrain exposed to a teratogen by cDNA microarray analysis.
Dheen ST; Hao AJ; Fu J; Gopalakrishnakone P; Ling EA
17455147 **Histol Histopathol** 22(7):729-42 2007
CD34 human hematopoietic progenitor cell line, MUTZ-3, differentiates into functional osteoclasts.
Ciraci E; Barisani D; Parafioriti A; Formisano G; Arancia G; Bottazzo G; Berardi AC
17533051 **Exp Hematol** 35(6):967-77 2007

Fig. 1. Compares the results of a relevance ranked citation search (top) with the same search ranked in reverse chronological order (bottom). The top three articles in each ranking are shown.

Optionally, the user can request that MiSearch use the results of the query itself to construct the profile. This results in a ranking where the citations sharing features with the largest number of other citations in the result set are ranked highly. In this view, citations that are most central to the topic rank highly while citations peripheral to the topic rank lower on the list. For example, in a search about a gene, citations where the gene is the major focus of the paper will be at the top of the query profile ranking while citations that only mention the gene in passing will rank lower on the list. Query profile mode is invoked by using 'query' or 'username query' as the username.

3.1 Evaluation

To assess the effectiveness of implicit relevance feedback, we use a cross validation approach. A training profile is constructed by sampling from the citations selected as relevant for viewing by a user. The test set consists of the remaining citations selected by this user. Typical results are shown in Figure 2. Increasing the number of citations in the training set

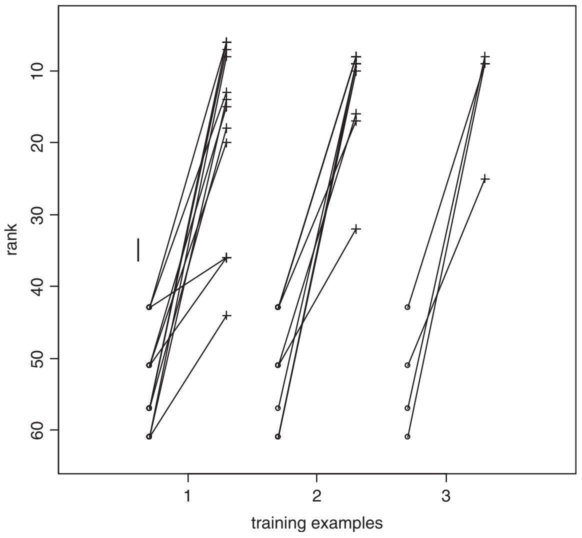


Fig. 2. Shown in the figure are the ranking for a representative search. The query 'Xist Tsix' that returns 66 articles in PubMed. The user selected four articles from this list related to epigenetic regulation of X inactivation. Leave one out cross validation of the relevance ranks were computed for training samples containing 1, 2 or 3 of the citations in the user's profile. The circles on the left show where each article appeared in the PubMed/Entrez ranking. The + on the right show the ranking of each article based on the MiSearch algorithm.

progressively improves the ranking of the test citations. In Supplementary Data, we compare the performance of MiSearch to relevance feedback using the Entrez 'related articles'/feature. The improved results in cross validation demonstrate that users are consistent in the articles that they select for viewing and that these selections are an effective implicit source of relevance feedback yielding improved biomedical literature search performance.

4 DISCUSSION

Automated collection of implicit relevant feedback information gathered using a click-through mechanism improve bibliographic search performance. Users find this interface intuitive and easy to use. Relevance feedback is applied by simply rerunning a query periodically during normal browsing. We find that response time is a critical factor in user acceptance of a relevance feedback system. While more sophisticated algorithms for classification and ranking based on relevance feedback have been proposed, the likelihood ratios used in MiSearch are effective and easily implemented within an RDBMS. This avoids the need to move large data sets in and out of the database server and improves user response time.

We encountered a number of issues in implementing MiSearch. Optimizing the performance of the relevance feedback system to work with small numbers of events is important. In a typical biomedical literature search task, users often view fewer than a dozen articles.

Many author names are not unique. In the MiSearch formulation, such author names are not resolved, but are expected to occur with higher frequency in the reference corpus and thus provide less information in ranking articles.

Documents vary greatly in the number of authors, MeSH terms and substance names applied to them. It is thus necessary to rank articles based on variable number of terms in these domains. We attempt to avoid bias in the formulation of the scores and by using pseudo counts where zero term counts give zero scores.

The MiSearch ranking is necessarily dependent on the NLM indexing processing. We are developing ways to base retrieval on automatically scored name, substance and MeSH headings, so that we can process documents such as web pages or journal articles that are not indexed by NLM.

Response time is an issue, particularly with very large result sets. The major performance bottleneck is that the system needs to calculate usage frequencies for every term appearing in every document in the result set. This is done on the fly so that rankings reflect the user's most recent search and retrieval behavior, but the reference term frequencies are pre-computed for all of PubMed. This is a compromise. For the task of ranking documents a user is likely to select, the reference corpus would ideally be the collection of documents that the users decided not to view among the citations that their queries had retrieve from Entrez. Implementing this would, however, be computationally intensive.

ACKNOWLEDGEMENTS

We thank Dr Angel W. Lee for her advice and comments during the course of preparing this manuscript. This work was supported in part by grants R01 LM008106 and U54 DA021519 from NIH.

Conflict of Interest: none declared.

REFERENCES

Bernstam,E. (2001) MedlineQBE (Query-by-Example). *Proc. AMIA Symp.*, 47–51.

Ding,J. *et al.* (2006) PubMed Assistant: a biologist-friendly interface for enhanced PubMed search. *Bioinformatics*, **22**, 378–380.

Kelly,D. and Teevan,J. (2003) Implicit feedback for inferring user preference: a bibliography. *ACM SIGIR Forum*, **37**, 18–28.

Muin,M. *et al.* (2005) SLIM: an alternative Web interface for MEDLINE/PubMed searches – a preliminary study. *BMC Med. Inform. Decision Making*, **5**, 37.

Nelson,S.J. *et al.* (2004) The MeSH translation maintenance system: structure, interface design, and implementation. *Medinfo*, **11**, 67–69.

Rocchio,J.J.J. (1971) Relevance feedback in information retrieval. In *The Smart System-experiments in Automatic Document Processing*. Prentice Hall Inc., Englewood Cliffs, NJ, pp. 313–323.

Salton,G. (1971) Relevance feedback and the optimization of retrieval effectiveness. In *The Smart System-experiments in Automatic Document Processing*. Prentice-Hall Inc., Englewocd Cliffs, NJ, pp. 324–336.

Salton,G. and Buckley,C. (1990) Improving retrieval performance by relevance feedback. *J. Am. Soc. Inform. Sci. Technol.*, **41**, 288–97.

Spink,A. *et al.* (2000) Use of query reformulation and relevance feedback by excite users. *Internet Res. Electr. Networking Appl. Policy*, **10**, 317–328.