BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 22 2009, pages 3038–3039
doi:10.1093/bioinformatics/btp529

Data and text mining

Identifying related journals through log analysis

Zhiyong Lu*, Natalie Xie and W. John Wilbur
National Center for Biotechnology Information (NCBI), 8600 Rockville Pike, Bethesda, MD 20852, USA

Received on June 16, 2009; revised on August 12, 2009; accepted on August 31, 2009
Advance Access publication September 3, 2009
Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: With the explosion of biomedical literature and the evolution of online and open access, scientists are reading more articles from a wider variety of journals. Thus, the list of core journals relevant to their research may be less obvious and may often change over time. To help researchers quickly identify appropriate journals to read and publish in, we developed a web application for finding related journals based on the analysis of PubMed log data.

Availability: http://www.ncbi.nlm.nih.gov/IRET/Journals

Contact: luzh@ncbi.nlm.nih.gov

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

1 INTRODUCTION

As a common practice, most of the scientists maintain familiarity with a small number of core journals to keep pace with the state of the art. Such a list is typically developed through years of personal experience and is highly dependent on an individual's research interests. In addition, the content of such a list may change over time. A previous study has shown that scientists are reading more articles on average per year from a wider variety of journals: increasing from 13 journals in 1977 to 33 individual journals by 2005 (Tenopir, 2008). This is no surprise because similar types of articles are now published in a broader range of journals and journals are covering a wider variety of topics and publishing more articles beyond their defined scope. Not only does this make it difficult for researchers to select journals for reading, it also makes for increasing difficulty deciding in which journal(s) to publish their own work (Schuemie and Kors, 2008). Thus, our objective is to suggest for users a list of current important journals related to journals they already know, so that researchers—especially scholars who are not yet deeply familiar with an area of research (e.g. junior graduate students)—may improve scholarly productivity.

To the best of our knowledge, very few systems/studies attempt to help scientists find relevant journals. The National Library of Medicine's (NLM's) Journals database provides a functionality that allows users to browse journals by discipline via Subject Terms: a set of MeSH® headings designated for indexing MEDLINE® journals by subject (e.g. biochemistry). However, Journal Subject Terms are frequently inadequate to find relevant journals across discipline boundaries, thus it may fail to meet individual needs. For instance, although *Bioinformatics* and *Nucleic Acids Research* are two closely related journals, they are indexed by completely different Journal Subject Terms.

JANE (Journal/Author Name Estimator) is a web server previously developed to help (i) authors find appropriate journals and (ii) editors find potential reviewers (Schuemie and Kors, 2008). However, by design JANE finds related journals through the set of MEDLINE citations that share a similar context with the input text, thus a short textual input such as a journal title often cannot yield optimized results (e.g. several top returned journals are not closely related to *Bioinformatics* when the word is used as input). Similar ideas for finding related journals can be seen in eTBLAST (Errami *et al*., 2007).

This work is also related to the use of clickthrough data to mine associations between items using techniques like collaborative filtering (Adomavicius and Tuzhilin, 2005). In particular, this is similar to the research on developing recommendation systems for large-scale digital libraries (Smeaton and Callan, 2001).

2 FINDING JOURNALS OF INTEREST

2.1 Browsing through Journal Subject Terms

At our web server, users can search for journals by browsing the same set of Journal Subject Terms. The distinction lies in how the resulting journals are sorted once a specific Subject Term is clicked. In addition to displaying journals alphabetically—the default order in the NLM's Journals database—we also list them by popularity, a measure determined by a journal's past usage.

2.2 Finding related journals

Alternatively, users can enter a query in the search box. Currently, the user can search a journal by its name, abbreviation or ISSN. The web server will return the bibliographic information of the requested journal such as its Publisher. Furthermore, there is a hyperlink called Related Journals. When clicked, it will display the 20 most related journals found by our approach (see below).

3 IMPLEMENTATION

We collected one month's (March, 2008) worth of the PubMed logs, which include a total of 8 million user sessions (after removing robot sessions) and 51 million *citation retrievals*. A citation retrieval is a specific MEDLINE record being clicked to display its corresponding bibliographic information and abstract text.

For each of the retrieval, we replaced it with its corresponding journal title in the dataset. A total of 15 827 journals (more journals

*To whom correspondence should be addressed.

than what is currently indexed in the Journals database) were found in the 8 million user sessions. The *usage* (number of times a particular journal was accessed) differs significantly between journals: some 1010 journals were heavily retrieved (over 10 000 times), while over 8000 journals were rarely viewed (less than 100 times).

Our calculation of related journals is based on the existence of a set of user sessions $\{s_i\}_{i=1}^{N}$ where each user session s_i consists of a set $\{d_j^i\}_{j=1}^{n_i}$ of citation retrievals in the form of MEDLINE records that were examined by the user during that session. If A represents a journal, we will denote by $t_A(s_i)$ the number of click through events that represent articles from journal A. We set

$$T_A = \sum_{i=1}^{N} t_A\,(s_i\,)$$

We may then estimate the probability of transitioning from an article in journal A to an article in journal B as

$$p\,(B|A\,) = \sum_{i=1}^{N} \left(\frac{t_A\,(s_i\,)}{T_A}\right)\left(\frac{t_B\,(s_i\,)}{n_i-1}\right)\left(\frac{n_i-1}{n_i}\right) = \sum_{i=1}^{N} \left(\frac{t_A\,(s_i\,)t_B\,(s_i\,)}{T_A n_i}\right)$$

Here the factor $\frac{t_A(s_i)}{t_A}$ represents the probability that a user looking at a document from journal A is actually looking at a document in session s_i. The factor $\frac{t_B(s_i)}{(n_i-1)}$ represents the probability that the next document (among the n_i-1 other documents represented in the session) that the user looks at will be a document from journal B. Finally, the factor $\frac{n_i-1}{n_i}$ represents the probability that the current record from journal A is not the last click through in the session. In this computation, we have assumed a random order to the clicked records making up the session, as we do not believe the order itself is important. In support of this assumption we remind the reader that the PubMed search engine retrieves documents in reverse time order of their entry into the database and there is a strong tendency for a user to click on them in that same order. Also note that in this computation, in addition to journal popularity, we have also taken its *topicality* (the relation of a journal to a user need) into consideration.

4 USER STUDY AND EVALUTION

To gain an understanding of how users respond to the new ranking of journals listed under the same Journal Subject Term (Section 2.1) and how accurately our approach can identify a customized list of related journals (Section 2.2), we conducted a user study by asking users to answer specific questions with regard to journals and their research interests (details in the Supplementary Material).

A total of 29 participants were recruited via email and personal contacts. They are comprised of graduate students, postdoctoral fellows or faculty members from a wide variety of biomedical subfields (e.g. genetics). They were first asked to identify their research field(s) and one or more journals they access the most. Next, they were asked to compare two different sorting strategies on the journal list indexed by Journal Subject Terms. One is based on usage (popularity), whereas the other is based on the alphabetic order. Finally, we asked them to evaluate the quality and usefulness of our computed journal list based on the journal they access the most. Specifically, participants were asked to identify irrelevant journals as well as other relevant journals missing from our suggestion list. As a result, both recall and precision can be computed for our journal suggestion lists.

Survey results are: first, for the list retrieved by a Subject Term, all users favored results ranked by usage over the alphabetic order, suggesting that popularity is an important factor in users' choice of favorite journals. Second, for the list retrieved by relatedness, the recall and precision are 0.893 (±0.102) and 0.910 (±0.086), respectively, suggesting that our computed journal list is closely related to a user's information needs. Third, all except one user favored our computed list over the list indexed by Subject Terms. Finally, with regard to the usefulness of such a list of related journals, >40% of our users (12/29) reported that they found at least one journal from the suggestion list to be important to their research but absent from their current checklist. All except two participants agreed that such a list is helpful, especially for scholars who are not yet deeply familiar with an area of science.

5 CONCLUSIONS AND DISCUSSION

We provide a web application for finding appropriate journals for researchers to read and publish in. Its unique feature of accurately identifying related journals is beyond the functionality of the NLM's Journals database and is complementary to other text mining tools such as JANE and eTBLAST.

Despite high recall and precision, our system failed to satisfy some researchers in the reported user study. No list of 20 journals could be guaranteed to include all important journals in an area. We also observed that some journals like *Nature* were repeatedly present in our results because of their popularity, but they were not always favored (because of their diverse content).

The web site is freely accessible and will be regularly updated. Part of our system has become available in the NLM's Journals database. The clickthrough data used in this research will be made available upon request after data anonymization, aggregation and transformation in accordance with proper user privacy protection.

ACKNOWLEDGEMENTS

We would like to thank all the participants for their contribution to the user survey.

Funding: Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of Interest: none declared.

REFERENCES

Adomavicius,G. and Tuzhilin,A. (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.*, **17**, 2005.

Errami,M. e*t al.* (2007) eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic Acids Res.*, **35**, 2007.

Schuemie,M.J. and Kors,J.A. (2008) Jane: suggesting journals, finding experts. *Bioinformatics*, **24**, 727.

Smeaton,A. and Callan,J. (2001) Joint DELOS-NSF workshop on personalisation and recommender systems in digital libraries. *ACM SIGIR Forum*, **35**, 7–11.

Tenopir,C. (2008) Online Databases: Are E-Journals Good for Sciences. Available at http://www.libraryjournal.com/article/CA6606485.html (last accessed date September 15, 2009).