MeSH term explosion and author rank improve expert recommendations

Danielle H. Lee, MS[1] and Titus Schleyer, DMD PhD,[2]

[1]School of Information Sciences; [2]Center for Dental Informatics, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA, USA

Abstract

Information overload is an often-cited phenomenon that reduces the productivity, efficiency and efficacy of scientists. One challenge for scientists is to find appropriate collaborators in their research. The literature describes various solutions to the problem of expertise location, but most current approaches do not appear to be very suitable for expert recommendations in biomedical research. In this study, we present the development and initial evaluation of a vector space model-based algorithm to calculate researcher similarity using four inputs: 1) MeSH terms of publications; 2) MeSH terms and author rank; 3) exploded MeSH terms; and 4) exploded MeSH terms and author rank. We developed and evaluated the algorithm using a data set of 17,525 authors and their 22,542 papers. On average, our algorithms correctly predicted 2.5 of the top 5/10 coauthors of individual scientists. Exploded MeSH and author rank outperformed all other algorithms in accuracy, followed closely by MeSH and author rank. Our results show that the accuracy of MeSH term-based matching can be enhanced with other metadata such as author rank.

INTRODUCTION

Information overload is an often-cited phenomenon that reduces the productivity, efficiency and efficacy of scientists. The volume of relevant information and resources, such as MEDLINE citations; gene sequences; tools and methods; and funding opportunities, is growing rapidly, often at an exponential rate. Electronic storage and transmission increase accessibility, but researchers typically retrieve most information they need on demand by actively searching for it. As the recently updated Long Range Plan of the National Library of Medicine points out, this creates a problem biomedical discovery:

"Millions of individuals … now retrieve terabytes of … health information and scientific data from NLM databases and services every day. … But most users … rely on a simple question and answer mode of querying … . Many important discoveries may never be realized because of this query method. … Enhancements that improve automated assistance to facilitate discoveries are badly needed." (NLM Board of Regents, 2006)

Faced with an ever-growing supply of information, researchers must invest increasing effort and time in routine information management, or risk missing relevant material and opportunities to advance their work. This is a particularly serious problem for researchers who are junior, engage in inter- and multi-disciplinary work, or lack a well-developed professional network. Therefore, there is a critical need to develop more effective and efficient ways of distributing, as opposed to producing, knowledge (Houghton, Steele et al. 2004).

Initiatives such as the NIH Clinical and Translational Science Awards and the Research Networking Program demonstrate the importance of developing informatics approaches to address information overload and improve information distribution within the biomedical research. Such approaches are increasingly developed in the emerging field of research informatics, for which AMIA recently launched dedicated conferences (the AMIA Summits on Translational Bioinformatics and Clinical Research Informatics).

We describe the development and formative evaluation of an algorithm to recommend scientists with similar research interests to each other. While the algorithm is generic and can compute the similarity of any pair of appropriately tagged information objects, we chose people because we could validate the performance of our algorithm against of the meaningful and easily obtainable gold standard of co-author relationships. We discuss how we used a vector space model (VSM) (Liu 2009) to calculate researcher similarity based on four approaches: 1) MeSH terms of publications; 2) MeSH terms and author rank; 3) exploded MeSH terms; and 4) exploded MeSH terms and author rank. We then describe how we evaluated the algorithm on a data set of 17,525 authors and their 22,542 papers.

RELATED WORK

Problem-solving, whether in industry or academia, is often a collaborative activity. Therefore, much research

in computer-supported cooperative work has focused on expertise location, i.e. determining "who knows what" and "who knows who knows what" within organizations (Wellman 2001). We briefly review selected approaches to expertise location, all of which either use a content- or social network-based approach, or a combination of the two.

ReferralWeb (Kautz, Selman et al. 1997) was an early attempt to locate experts using social networks. This research prototype used a social network graph in order to allow users to find short referral chains to suggested experts quickly. Social networks and expertise profiles were constructed by mining publicly available Web documents. The system perceived pairs of users co-appearing on a Web page as socially connected, and inferred personal expertise through Webpages that mentioned people and topics together. This approach, however, holds high uncertainty in depicting social networks and expertise. It is also not sure how well it would apply to organizations in which expertise and social connections are often represented differently.

SmallBlue (Lin, Cao et al. 2009) is an internal IBM system that helps users find experts for a certain topic. It is both content- and social network-based, and visualizes the social networks of experts when queried for a specific topic. The system employs private emails and chat logs to determine expertise and social connections. Even though SmallBlue users grant the system explicit access to their personal communications logs, privacy issues may reduce its applicability in other settings, especially academia.

Yang and Chen (Yang and Chen 2008) describe an educational P2P (peer-to-peer) system at a Taiwanese university. When queried for a term, it recommends items posted by users with the highest expertise scores and who are most preferred by the target user. In order to function, human experts should assess each user's expertise and users have to rate other users explicitly. That means the system needs significant human intervention that is unlikely to be sustained, especially for general-purpose systems.

The Expertise Oriented Search (EOS) system (Li, Tang et al. 2007) is designed to allow users to identify expertise and explore social associations of researchers in computer science. To do so, the system draws on a researcher's 20 most relevant Web pages retrieved from Google and a publication list as obtained from the Digital Bibliography and Library Project, and Citeseer, respectively. Topic relevance is propagated through social connections, assuming that a person's expertise diffuses through interactions in social networks. Both original topical expertise and propagated relevance values are taken into account during searches.

McDonald (McDonald 2003) introduced a system to recommend experts within a software company. The recommendation algorithm integrates two kinds of social networks: work context- and sociability-based. The social networks are constructed partially through user preferences, and partially by researchers using various ethnographic methods. An evaluation did not identify one type of network as superior over the other, but suggested that there was a trade-off in recommendations when considering only expertise or social connections, respectively. The social networks in the system were created entirely through manual means, making the approach hard to use in other contexts.

Pavlov and Ichise (Pavlov and Ichise 2007) analyzed the structure of social networks to predict collaborations at a Japanese science institution. They used graph theory to build feature vectors for each expert dyad and applied four machine learning methods (support vector machines, two decision trees and boosting) to predict collaborations. The two decision tree techniques outperformed when precision and recall were combined, and all algorithms were better than the random (control) approach.

Bedrick and Sitting's (Bedrick and Sittig 2008) Medline Publication (MP) Facebook application is one system described for biomedical research that relies entirely on content for expert recommendations. MP models expertise using MeSH terms drawn from publications. It recommends potential collaborators by comparing the angle of small expertise vectors calculated using singular value decomposition.

As this brief review shows, many expert recommendation systems integrate content-based with social recommendations. Social networks are either inferred through computation or defined by users themselves. Inferred social networks tend to be subject to a large degree of uncertainty (Backstrom, Huttenlocher et al. 2006). On the other hand, it is hard to expect users to specify their social connections in a real-world context. In addition, many of the described systems suffer from limitations that restrict their ability to recommend experts in biomedical research. In this study, we combined MeSH term matching with other metadata, in our case author rank, to generate recommendations for "similar" people in biomedical research. In the following section, we explain our recommendation algorithm and the data we used in our evaluation.

RECOMMENDATION ALGORITHM AND EXPERIMENTAL DATA SET

Our recommendation algorithm is based on the vector space model (VSM), one of the most commonly used approaches in information retrieval (Liu 2009). To

calculate the similarity of two documents, first all terms (all words excluding stop-words) in each document are counted. Then, document similarity is determined by the degree to which the same words appear in either document, using a Cosine correlation.

In this study, we substitute authors for documents and their papers' MeSH terms for document terms. We evaluate four types of inputs for our algorithm: 1) MeSH terms; 2) MeSH terms and author rank, i.e. the position of the author in the author list; 3) exploded MeSH terms; and 4) exploded MeSH terms and author rank. The first two approaches are naïve while the latter are extended techniques designed to increase the scope of the similarity comparison.

The MeSH term-based approach is the simplest because it only considers the collective MeSH terms assigned to each author's publications. To calculate the Cosine similarity of two authors (a_i and a_j), the Term Frequency and Inverse Document Frequency (TF/IDF) of their MeSH term collections are calculated as shown in Equations 1 and 2.

$$w_{in} = tf_{in} \times \mathrm{idf}_n \qquad \textbf{eq. 1}$$

$$Cosine(a_i, a_j) = \frac{\sum_{n=1}^{|V|} w_{in} \times w_{jn}}{\sqrt{\sum_{n=1}^{|V|} w_{in}^2 \times \sum_{n=1}^{|V|} w_{jn}^2}} \qquad \textbf{eq. 2}$$

In order to determine author similarity, we first calculate TF/IDF of each MeSH term that an author has published on (Equation 1). Variable w_{in} denotes the TF/IDF values of a MeSH term n in author a_i's publications. It is the product of term frequency (tf_{in}) and inverse document frequency (idf_n) (Liu, 2007). Term frequency tf_{in} measures how many times a term n appears in the author a_i's publications (Table 1). Our algorithm design assumes that the higher the term frequency, the higher the presumed expertise of the author on the subject.

Table 1. Example of Authors' Term Frequency

Term	Author1	Author2	Author3	Author4
diabetes	115	53	11	20
ileum	10	1	0	12
neoplasm	0	8	38	2

However, term frequency alone is insufficient to calculate similarity because terms that occur frequently *across* many papers do not distinguish authors very well from each other. Therefore, we apply inverse document frequency (idf_n) to compensate for this limitation. *idf* emphasizes terms which occur less frequently across documents, and are, as a result, more informative and discriminative.

We use the TF/IDF values of pairs of authors (a_i and a_j) to calculate their similarity. The variable V is a union set of MeSH terms that a_i and a_j have. The Cosine similarity is computed using the TF/IDF values of all terms of both authors.

In the second approach, we combine MeSH terms with authorship rank (Equations 3 and 4) because we hypothesize that author rank is correlated with expertise. Typically, the first author is considered the main expert on the topic of the paper. In this project, we make the simplified assumption that all authors' expertise on the topic of a paper is proportional to their position in the author list. While this assumption may not hold in all cases (esp. for papers authored by trainees and their advisors), it simplifies algorithm design.

$$o_{in} = \sum_{m=1}^{M} (ta_m - (ao_{im} - 1))/ta_m \qquad \textbf{eq. 3}$$

$$Cosine(a_i, a_j) = \frac{\sum_{n=1}^{|V|} w_{in} o_{in} \times w_{jn} o_{jn}}{\sqrt{\sum_{n=1}^{|V|} (w_{in} o_{in})^2 \times \sum_{n=1}^{|V|} (w_{jn} o_{jn})^2}} \qquad \textbf{eq. 4}$$

Variable ao_{im} denotes the weight of author a_i's author rank for MeSH term n. M is the set of his publications that the corresponding MeSH term is assigned to. ta_m is the total number of authors on the paper and ao_{im} is the rank of author a_i. For example, in eq. 5, Author1 is the 1st of three authors and the 4th of 11 authors on two papers indexed with the Term A, yielding a value of 1.73 for o_{1A}. o_{1A} thus provides the weighted sum of Author1's expertise on Term A.

$$o_{1A} = \frac{3 - (1 - 1)}{3} + \frac{11 - (4 - 1)}{11} \qquad \textbf{eq. 5}$$

We chose our third approach (exploded MeSH terms) because the fine-grained nature of the MeSH hierarchy (as of 2009, 50,956 terms in 11 hierarchical levels) may make it difficult to determine the true semantic similarity of papers. Two very closely related papers might be indexed with sibling terms, but would not be considered similar using the first two algorithms we have described. Therefore, our third approach explodes source MeSH terms and only considers children at the leaf level. We excluded MeSH terms at the top level because we considered them to be too general to be discriminative. We calculated author similarity using TF/IDF as described in our first approach.

In the last approach, we combine exploded MeSH terms and author rank. We calculate the TF/IDF values for all leaf terms and multiply these values with the weights derived from author rank. We evaluate the performance of our approaches by comparing actual co-author relationships (gold standard) with those predicted by our algorithms. To do so, we constructed a data set using the snowball method starting with 200 randomly sampled seed authors in the University of Pittsburgh's Faculty Research Interests Project System (Friedman, Winnick et al. 2000). Snowball sampling is considered superior to other approaches such as node or link sampling since the latter techniques are likely to include many isolated pairs (Ahn, Han et al. 2007). We expanded our sample by including all co-authors and the co-authors' co-authors through breadth-first search. Collexis Holdings, Inc., Columbia, SC, provided the data set which was fully disambiguated, i.e. authors and their relationships were unambiguously specified using an approach similar to that described by Torvik et al. (Torvik, Weeber et al. 2005). The data set included full citation and author relationship information. We added MeSH terms for publications directly from PubMed. Table 2 describes the sample.

Table 2. Experimental Data Set

No. of authors	17,525
No. of publications	22,542
Avg. no. of papers per author	5.4
No. of papers that at least one MeSH term was assigned to	21,806
Avg. no. of MeSH terms per paper	22.9
Avg. no. of exploded MeSH terms per paper	114.8

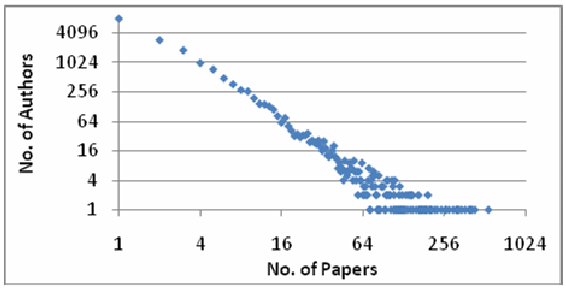

Figure 1. **Number of papers per author**

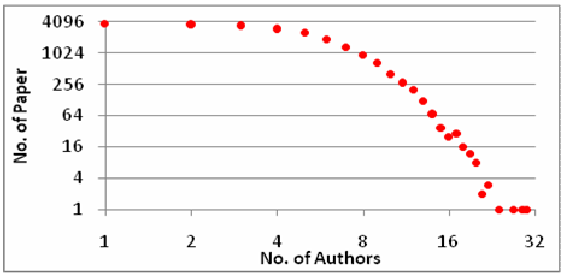

Figure 2. Number of authors per paper

Figures 1 and 2 illustrate the number of papers per author and the number of co-authors per paper. More than half of the authors (9,650 authors or 55.1%) have published more than one paper. Most papers (18,782 papers or 83.3%) have more than one author. This indicates that the data sets may have sufficient overlap among authors to be able to calculate author similarity. The mean number of MeSH terms per paper is 22.9 (σ = 10.9).

We evaluated the performance of our algorithms as follows. For each of 150 authors selected at random from our sample, our algorithms determined the five and 10 most similar authors (herein, Top 5/10 authors), regardless whether they co-authored papers or not. Then, we checked how many of the Top 5/10 authors actually did co-author a paper with the test author. We used a Friedman two-way ANOVA test to compare the mean difference between the number of correctly predicted co-authors. The difference was considered statistically significant at a p value of 0.01. In a second analysis, we calculated how many papers correctly identified co-authors wrote together. A higher number was considered indicative of a closer working relationship, and thus a better recommendation. As described above, we used the Friedman two-way ANOVA test to compare mean differences for paper averages.

EXPERIMENTAL EVALUATION

Table 3 shows the number of actual co-authors in the Top 5/10 evaluation categories predicted by the four algorithms. In each evaluation category, the algorithms predicted an average of approximately 2.5 authors correctly. When analyzed using the Friedman two-way ANOVA, prediction accuracy between any two pairs was significantly different (χ^2 = 108.44, $p < .001$ for Top 5, χ^2 = 141.39, $p < .001$ for Top 10). Exploded MeSH and author rank outperformed all other algorithms in both Top 5/10, followed closely by MeSH and author rank.

Table 3. Average number of correctly predicted co-authors in Top 5/10 co-authors

	Top 5	**Top 10**
MeSH	2.08	1.90
MeSH & author rank	2.72	2.72
Exploded MeSH	2.38	2.40
Exploded MeSH & author rank	2.82	2.98

Table 4. Average number of co-authored papers in Top 5/10

	Top 5	**Top 10**
MeSH	10.07	10.38
MeSH & author rank	12.56	12.37
Exploded MeSH	10.94	11.54
Exploded MeSH & author rank	13.03	12.57

Table 4 shows how many papers correctly identified co-authors wrote together. For this evaluation criterion, all four approaches performed with a statistically significant difference (χ^2 = 11.70, p = .008 for Top 5, χ^2 = 12.39, p = .006 for Top 10). Both MeSH and exploded MeSH, combined with author rank, performed best.

CONCLUSION AND DISCUSSION

This paper introduced a novel expert recommendation algorithm that combined naïve and extended MeSH term matching with author rank, and evaluated its performance in matching experts against the gold standard of co-authorship. We found that the hybrid approach of exploded MeSH terms and author rank performed best, followed by the combination of MeSH terms and author rank. It therefore appears that adding relevant metadata such as author rank can improve the performance of expert recommendation algorithms.

It should be noted that this study only represents an initial attempt to improve the performance of term-based recommendation algorithms with other metadata. The results we obtained should be verified and generalized with other, possibly larger, data sets.

One limitation of our study was that we focused solely on author similarity. Correctly recommending co-authors to a target user has little practical value. However, this limitation allowed us to exploit a significant methodological strength: validation against an excellent gold standard, i.e. actual co-author relationships. When evaluating our algorithms in field studies, we will omit coauthors from the recommendations, yielding potentially useful recommendations of "similar people." A second limitation was the fact that we attributed expertise through author rank in a simplistic way that does not take the varied contributions of authorship into account.

In future work, we plan to refine our algorithms by adding other metadata, for instance publication types. In addition, we intend to study recommending *complementary*, as opposed to *similar*, people using algorithms such as those developed by Swanson and Smalheiser (Swanson and Smalheiser 1997). Last, we need to find ways to reduce the size of the vector space using latent semantic indexing or other clustering methods. As other researchers have pointed out (Bedrick and Sittig 2008), naïve vector calculations consume a lot of time and resources. Lastly, we will investigate how to recommend

ACKNOWLEDGEMENTS

We thank Collexis Holdings, Inc., for providing the data set, and gratefully acknowledge the support of the National Center for Research Resources for this project (grant number UL1 RR024153).

REFERENCES

[1] Ahn, Y.-Y., S. Han, et al. (2007) Analysis of topological characteristics of huge online social networking services. *Procs. of the 16th Intl. Conf. on World Wide Web*, Banff, Alberta, Canada.

[2] Backstrom, L., D. Huttenlocher, et al. (2006) Group formation in large social networks: membership, growth, and evolution. *Procs. of the 12th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, Philadelphia, PA, USA.

[3] Bedrick, S. and D. Sittig (2008) A scientific collaboration tool built on the facebook platform. *Procs. of AMIA 2008 Ann. Symp.*, Washington, DC, USA.

[4] Friedman, P., B. Winnick, et al. (2000) Development of a MeSH-based index of faculty research interests. *Procs. of AMIA 2000 Ann. Symp*.

[5] Houghton, J. W., C. Steele, et al. (2004). Research practices and scholarly communication in the digital environment. *Learned Publishing* 17: 231-249.

[6] Kautz, H., B. Selman, et al. (1997). Referral Web: combining social networks and collaborative filtering. *Commun. ACM* 40(3): 63-65.

[7] Li, J., J. Tang, et al. (2007) EOS: expertise oriented search using social networks. *Procs. of the 16th Intl. Conf. on World Wide Web*, Banff, Alberta, Canada.

[8] Lin, C.-Y., N. Cao, et al. (2009). SmallBlue: Social Network Analysis for Expertise Search and Collective Intelligence. *IEEE Intl. Conf. on Data Engineering*.

[9] Liu, B. (2009). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*, Springer.

[10] McDonald, D. W. (2003) Recommending collaboration with social networks: a comparative evaluation. *Procs. of the SIGCHI Conf. on Human Factors in Computing Systems*, Ft. Lauderdale, Florida, USA.

[11] Pavlov, M. and R. Ichise (2007) Finding Experts by Link Prediction in Co-authorship Networks. *Procs. of the Workshop on Finding Experts on the Web with Semantics (FEWS2007) at ISWC/ASWC2007, Busan, South Korea*.

[12] Torvik, V. I., M. Weeber, et al. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation: Research Articles. *J. Am. Soc. Inf. Sci. Technol.* 56(2): 140-158.

[13] Wellman, B. (2001). Computer Networks As Social Networks. *Science* 293(5537): 2031-2034.

[14] Yang, S. J. H. and I. Y. L. Chen (2008). A social network-based system for supporting interactive collaboration in knowledge sharing over peer-to-peer network. *Int. J. Hum.-Comput. Stud.* 66(1): 36-50.

[15] Swanson D. R. and Smalheiser N. R (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif. Intell.* 91(2): 183-203.