Expert Recommendation for Knowledge Management in Academia

Tamara Heck, Oliver Hanraths, Wolfgang G. Stock
Heinrich-Heine-University, Dept. for Information Science
D-40225 Düsseldorf, Germany
{Tamara.Heck;Oliver.Hanraths}@hhu.de, Stock@phil.hhu.de

ABSTRACT
Recommendation systems are not only important in e-commerce, but in academia as well: They support scientists in finding relevant literature and also potential collaboration partners. It is essential that such a recommendation system proposes the most relevant people. Scientometric similarity measurements like co-citation and bibliographic coupling analysis have proved to give a good representation of research activities and hence it can be said that they put authors with similar research together and detect possible collaborations. Our aim is to implement a recommendation system for a target author who searches for collaboration colleagues. The research question is: 1) Can we propose a relevant author cluster for a target scientist? Furthermore we try to apply user data from the social bookmarking system CiteULike. The second research question is: 2) Is this user-based data also relevant for our target scientist and does it recommend different results? Our first outcomes of this work in progress are evaluated by our target authors.

Keywords
Expert recommendation, social bookmarking, Similarity measurement, Knowledge management, Academia

INTRODUCTION: EXPERT RECOMMENDATION

Knowledge Management in Academia

An important task for knowledge management in academic settings and in knowledge-intensive companies is to find the "right" people who can work together to solve a scientific or technological problem successfully. Exemplarily, we will list some situations in which expert recommendations are very useful:

- compilation of a (formal) working group in a large university department or company,
- compilation of researchers and appropriate co-authors for preparing a project proposal for a research grant,
- forming a Community of Practice, independent from the affiliation with the institutions,
- accosting colleagues in preparation of a congress or for contributions to a handbook.

Starting point of expert recommendation systems are both, topics (e.g., scientific-technological tasks) and available experts (e.g., searching other currently unknown experts fitting best to "our" staff members). It is very important for cooperation in science and technology that the reputation of the experts is proved. The reputation of an expert in science and technology grows with his amount of publications in peer-reviewed journals and with the citations (Cronin, 1984, p. 13). Multi-discipline information services which allow publication and citation counts are Web of Science (WoS) and Scopus. All comparative studies indicate, on the document and the author level, that both services show different results (e.g. Meho & Sugimoto, 2009). Our approach analyzes data from WoS and Scopus. Additionally, our system makes use of data from CiteULike (CUL), which is a social bookmarking service for academic literature (www.citeulike.org). So we can consider not only the authors' perspectives (by tracking their publications, references and citations), but also the perspectives of the readers (by tracking their bookmarks and tags).

Related Work

There are several approaches to construct expert recommendation systems, of which we will mention some examples: Au Yeung et al. (2009) try to detect experts in the non-academic bookmarking system Del.icio.us, defining an expert as someone who has high-quality documents in his bookmark collection. Chenthamarakshan et al. (2009), Petry et al. (2008), Reichling and Wulf (2009) and Yukawa et al. (2001) investigate in expert recommendation mainly for business institutions. Blazek (2007) focuses on expert recommendation sets of articles for a "Domain Novice Researcher", i.e. for new academics. Zanardi and Capra (2008), proposing a "Social Ranking", calculate similarity between users based on same tags and tag-pairs based on same bookmarks in CUL. The results show that user similarity improves accuracy whereas tag similarity improves coverage. Heck and Peters (2010) use three social bookmarking systems (CUL, Connotea, Bibsonomy) to recommend researchers, who are unknown to the target

ASIST 2011, October 9-13, 2011, New Orleans, LA, USA.

researcher and could be potential collaboration partners to build communities of practice. The approach of Cabanac (2010) is similar. He concentrates on user similarity networks and relevant articles. He uses the concepts of Ben Jabeur et al. (2010) to build a social network for recommending relevant literature. Nocera and Ursino (in press), recommending similar users and resources, set their focus on "social folksonomy" while using information about user friendships and semantic information of tags (see also De Meo et al. 2011).

EXPERIMENTAL METHODS

Used Algorithm and Dataset

According to van Eck and Waltman (2009) the most popular similarity measures are the association strength, the cosine, the inclusion index and the Jaccard index. With the appearance of bookmarking and collaboration services, several other algorithms were developed. As variants using collaborative filtering methods (Goldberg et al., 1992) they differ in combination of the considered relations between users, items and tags and the used weights. In our approach we make use of the cosine. Our own experience (Heck, in press) and other papers (e.g. Rorvig, 1999) show that cosine works well. But in later project steps we also want to try Dice and Jaccard-Sneath.

What is new for our system is the attempt to compare and combine data from WoS, Scopus, and CUL, which makes our approach particularly valuable for academic knowledge management. We concentrate on recommending similar researcher unknown to our scientists, i.e. we don't consider co-authorship. Our starting point is a single researcher, not a topic: We take the author's publications from 2006-2011 to base recommendations on current research interest. Based upon the data from WoS, Scopus and CUL, we sort all related authors of a target author by similarity and cut the list at *k* names. For those *k* names we calculate bibliographic coupling, author co-citation as well as user and tag similarity. The result in each case is a *k*k* similarity matrix represented as a single-link cluster. WoS is used to analyze bibliographic coupling. We aggregate the data from the document level to the author level. Bibliographic coupling (**BC**) of authors means that two authors A1 and A2 are linked if they cite the same references. Author co-citation (**ACC**) means that two authors A1 and A2 are linked if their works are cited in the same article. To mine ACC data it is not possible to work with WoS because in the references section of a bibliographic entry only the first author of the cited documents is named and not, what is needed, all authors (Zhao & Strotmann, 2011). Therefore we use Scopus. If we work with the cosine, the similarity of authors between A1 and A2 is $SIM = g / (ab)^{1/2}$, where *g* is either the number of references A1 and A2 have in common (BC) or the number of documents citing A1 and A2 (ACC), and *a* respectively *b* is the number of references (BC) or citing documents (ACC) of A1 respectively A2. The method in CUL is based on collaborative filtering (e.g., Goldberg et al., 1992). All tags which users assigned to the target author's articles are considered to find the most similar documents. We assume that the authors of the documents, which have at least *n* tags in common with the target author's documents, are similar to him and frame our database. Author similarity is measured either based on common users or on common tags.

In all three databases problems arise, which we will point out briefly as the recommendation results highly depend on the source dataset. In Scopus there are discrepancies in the bibliographic data: e.g. title and authors may be complete in one reference list, but incomplete in another. In our dataset, several co-authors were missed and couldn't be considered for ACC. The completeness highly varies: In a random sample, where the co-citation dataset is adjusted with data of the Scopus website, five of 14 authors have a complete coverage, three of them have coverage between 70 and 90 %, six authors a coverage under 70 %. There is also the problem of name homonymy. The author Id-number in Scopus is supportive for identification, but it may fail if two or more homonymic authors are allocated to the same research field. In WoS, where there isn't any author-id, we checked the authors' document lists and if necessary correct them based on the subject area. Additionally in CUL users may misspell names, which had to be rechecked.

Evaluation

Our method gives us four sets of potential relevant authors, which we call clusters: One cluster is based on ACC in Scopus (**COCI**), one cluster is based on BC in WoS (**BICO**), one cluster is based on common users in CUL (**CULU**) and one cluster is based on common tags in CUL (**CULT**). Additionally we visualize the clusters, what we call graphs, and show the cosine similarity between all the authors of one cluster. The graphs will help the physicists to evaluate author similarity. The evaluation is divided in two parts: In part one the researcher has to rank the proposed similar authors according to their relevance. Therefore the ten top authors of all four clusters are listed in alphabetical order (co-authors eliminated). The interviewee should tell how important they are for his research (rating from not important (1) to very important (10)), with whom he would cooperate and which important people he misses. In part two our author has to evaluate the graphs (rating 1 to 10) according to relevance and the distribution of the authors.

RESULTS

Part 1. Cluster Evaluation

Six physicists evaluated their individual clusters. We will demonstrate first results on one example: Our researcher (physicist #1) published 8 articles (2006-2011). In general the cosine coefficient is much higher for ACC than for BC because some authors have a lot of references, which minimize similarity: The ACC similarity is between 0.82 and 0.04 whereas the BC similarity is between 0.12 and 0.01. Both measures show different results, i.e. diverse authors. Only one author is in both clusters. In CUL the similarity distribution between 0.87 (0.74 tag-based) and 0 (0.16 tag-based) is much wider. The results may be a hint to

modify the cosine for cluster combination. 20 authors are found twice, i.e. they are in the CUL cluster and either in the ACC cluster (19) or in the BC cluster (1). The great differences, found by almost all target authors, between the ACC and BC results struck some interviewees. It has to be analyzed if it depends on the different databases.

	Co-Citation	Bib. Coupling	CiteULike
Co-Citation	11 (3)	0	8
Bib. Coupling	0	4 (3)	1
CiteULike	8	1	15 (6)

Table 1. Number of located important authors (ranked by target author) based on different relations.
Numbers in brackets: authors located in only one source.

We took the ten top authors of each of the four clusters. Author duplicates and co-authors were eliminated, hitherto 30 top authors remained for physicist #1. From these 30 authors the interviewee knew 16 people (7 personally). Concerning the importance of these people, she assigned a 10 (very important) to 8, an average importance to 5 and no importance to one author (three authors couldn't be ranked). The interviewee would cooperate with 7 authors. As reasons for non-cooperation she named "direct competitors" and "low thematically overlap", two reasons which are also mentioned by the other interviewees. Beneath the 13 important authors on the list physicist #1 named 18 other authors, who are important for her, but aren't located under the 30 top ranked people. Eliminating 2 co-authors, 29 important authors were left, for whom we

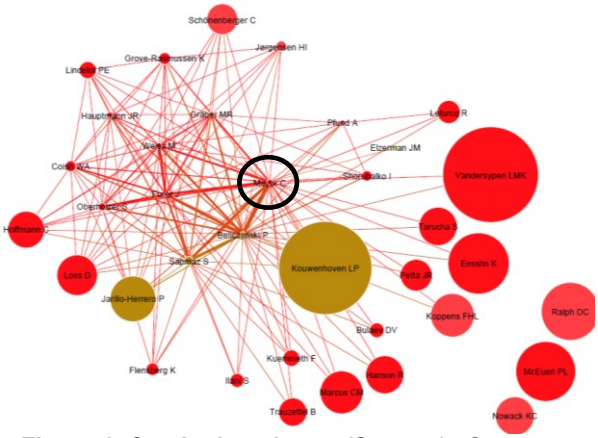

Figure 1. Co-citation cluster (Scopus) of a target author (circle), cosine threshold 0.06.

analyzed the results (Table 1). It is surprising that CUL provides the most authors and almost half of them are not located with ACC and BC. Similar results were found by another physicist, whereas results vary by the other interviewees, where BICO gives the most relevant people.

8 important authors of physicist #1 are in none of the datasets. Comparing the ranks for the 21 located authors, there are great differences. In general an author in one of the CUL datasets (CULU or CULT) has a higher rank. The reason for this might be the major number of authors found. Modifying the cosine coefficients and adding the results leads to a better ranking in ten cases, but for some authors to a worse rank. 9 important authors are ranked within the 60 top authors of the cumulated results. If we take the 15 top authors of each of the four datasets, we get 12 important authors. This result allows two conclusions: Either the other unknown authors don't belong to physicist's #1 research field and are thus unimportant for her, i.e. the similarity measurements don't show the most relevant authors. Or these authors are important (but yet unknown to our target scientist) and are potential collaboration partners. The more appropriate conclusion can be drawn when the target author sees the graphs and the location of the unknown authors in relation to her important researchers.

Part 2. Graph Evaluation

We visualized the clusters with graphs: The edges are seized according to the cosine weight, the nodes (authors) were first seized according to their number of citing documents of an author (Fig. 1) or assigned tags to his articles (Fig. 2). But as the node seize irritated our exemplary interview partner (physicist #1 supposed bigger nodes to show higher similarity), we changed this aspect by the other interviewees and adjusted node seize according to an author's number of relations to others authors. We set thresholds to have clear arrangements in the graphs for better evaluation. Furthermore in the CUL clusters we left out author-pairs with sim= 1 if these authors have only one tag in common (see Fig. 2) because this causes a biased image of the author community. Physicist #1 claimed CULT the best (relevance = 9), followed by the graph based

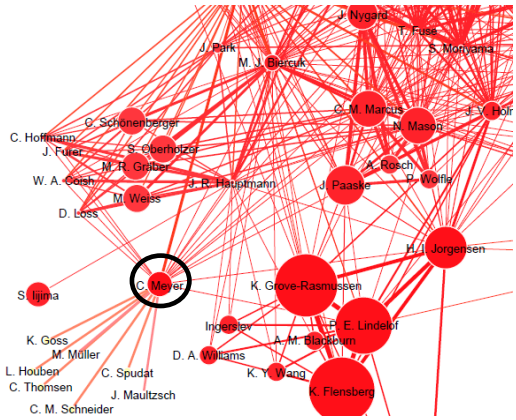

Figure 2. Extract of tag-based cluster (CiteULike) of a target author (circle), cosine interval 0.99-0.45.

on BICO (relevance = 7) and the COCI graph (relevance = 5-6). CULU was claimed the worst with a relevance of only 3-4. CULT best covers the physicist's research field and show distinctive mapping of author groups (Fig.2: more single communities visible). Physicist #1 claimed that the unknown authors of the list might be relevant if they are related to her important authors in the cluster. The average graph relevance (based on six target users) is:

BICO: 8.7 **CULT**: 5.25 **COCI**: 5.08 **CULU**: 2.13

Consider that only four authors had publications in CUL to be analyzed. Two authors claimed BICO and CULT to be very relevant and proposed to combine these two to get all

important authors and relevant research communities. Physicist #1 would also combine her CULT graph (Fig. 2) with COCI (Fig. 1). In BICO and COCI she missed important authors. A combined cluster could help her to find researcher groups, partners for cooperation and it would be supportive to intensify relationships among colleagues. Additionally two interviewees would prefer bigger clusters like the CUL graphs because they show more unknown and possible relevant people. Looking at the clusters all physicists recollected important colleagues, who didn't come to their mind first, which they found very helpful. However an important factor for them is a clear cluster arrangement. A problem which may concern CUL clusters is the sparse dataset, i.e. if only few tags were assigned to an author's articles or few users bookmarked them, the cluster cannot show high distinguishable communities. That was the case with one author, who claimed CULU and CULT worse than COCI and BICO.

DISCUSSION

In our project we analyzed expert recommendation based on different author relations in three databases as a new approach to recommend relevant experts in academia. We combined two scientometric approaches (ACC and BC) with collaborative filtering methods. First results show that the combination of different methods leads to the best results. Similarity based on data of a social bookmarking system may complement ACC and BC. The interviewees approved this assumption with the cluster relevance ranking. They and the other researchers, also in former studies (Heck & Peters, 2010), confirm that there is a need for author recommendation. Further study will concentrate on the adequate similarity coefficient and the attempt to combine results of different databases as well as discuss social network analysis and graph construction.

ACKNOWLEDGMENTS

Tamara Heck and Oliver Hanraths are supported by the *Strategische Forschungsfonds* of the Heinrich-Heine-University Düsseldorf. Thanks to Isabella Peters and Stefanie Haustein for valuable discussions and the physicists for evaluation.

REFERENCES

Au Yeung, C. M., Noll, M., Gibbins, N., Meinel, C., & Shadbolt, N. (2009). On measuring expertise in collaborative tagging systems. *WebSci '09*. Athens.

Ben Jabeur, L., Tamine, L., & Boughanem, M. (2010). A social model for literature access: towards a weighted social network of authors. *RIAO '10* (pp. 32-39). Paris.

Blazek, R. (2007). Author-Statement Citation Analysis Applied as a Recommender System to Support Non-Domain-Expert Academic Research. Nova Southeastern Univ. http://gradworks.umi.com/32/78/3278195.html.

Cabanac, G. (2010). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics*, 87 (3), 597-620.

Chenthamarakshan, V., Dey, K., Hu, J., Mojsilovic, A., Riddle, W. & Sindhwani, V. (2009). Leveraging social networks for corporate staffing and expert recommendation. *IBM Journal of Research and Development*, 53 (6). 11:1-11:10.

Cronin, B. (1984). *The Citation Process. The Role and Significance of Citations in Scientific Communication*. London, UK: Taylor Graham.

De Meo, P, Nocera, A., Terracina, G & Ursino, D (2011). Recommendation of similar users, resources and social networks in a Social Internetworking Scenario. *Information Sciences, 181* (7)*,* 1285–1305.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12), 61-70.

Heck, T. (in press). A Comparison of Different User-Similarity Measures as Basis for Research and Scientific Cooperation. *Issome '11*. Turku, Finland.

Heck, T., & Peters, I. (2010). Expert recommender systems: Establishing Communities of Practice based on social bookmarking systems. *I-Know '10* (pp. 458-464). Graz.

Meho, L. I., & Sugimoto, C. R. (2009). Assessing the scholarly impact of information studies. A tale of two citation databases – Scopus and Web of Science. *JASIST*, 60 (12), 2499-2508.

Nocera, A., & Ursino, D. (in press). An approach to provide a user of a "social folksonomy" with recommendations of similar users and potentially interesting resources. *Knowledge-Based Systems*,DOI:10.1016/j.knosys.2011.06.003.

Petry, H., Tedesco, P., Vieira, V., & Salgado, A. C. (2008). ICARE. A context-sensitive expert recommendation system. *ECAI'08* (pp. 53-58). Patras, Greece.

Reichling, T., & Wulf, V. (2009). Expert recommender systems in practice. Evaluating semi-automatic profile generation. *CHI '09* (pp. 59-68). New York: ACM.

Rorvig, M. (1999). Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. *JASIST*, 50 (8), 639-651.

Van Eck, N. J., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *JASIST*, 60 (8), 1635-1651.

Yukawa, T., Kasahara, K., Kita, T. & Kato, T. (2001). An Expert Recommendation System using Concept-based Relevance Discernment. *ICTAI '01* (pp. 257). Dallas.

Zanardi, V., & Capra, L. (2008). Social ranking: Uncovering relevant content using tag-based recommender systems. *RecSys '08* (pp. 51-58). New York: ACM.

Zhao, D., & Strotmann, A. (2011). Counting first, last, or all authors in citation analysis. Collaborative stem cell research field. *JASIST*, 62 (4), 654-676.