Combining Social Information for Academic Networking

Tamara Heck
Heinrich-Heine-University Duesseldorf
Department of Information Science
Universitaetsstrasse 1, Duesseldorf, D-40225
tamara.heck@hhu.de

ABSTRACT
Researchers in almost all scientific disciplines rely heavily on the collaboration of their colleagues. Throughout his or her career, any researcher will build up a social academic network consisting of people with similar scientific interests. A recommendation system could facilitate the process of identifying and finding the right colleagues, as well as pointing out possible new collaborators. As a researcher's reputation is of great importance, the social information gleaned from citations and reference data can be used to cluster similar researchers. Web services, such as social bookmarking systems, provide new functionalities and a greater variety of social information – if exploited correctly, these could lead to better recommendations. The following chapter describes, by way of example, one approach to recommendation for social networking in academia.

Author Keywords
Social information; scientific collaboration; author recommendation; folksonomy; collaborative filtering; author co-citation; bibliographic coupling

ACM Classification Keywords
H.2.8 **[Database Management]:** Database Applications – Scientific databases. H.3.3 **[Information Storage and Retrieval]:** Information Search and Retrieval – Information filtering. H.3.5 **[Information Storage and Retrieval]:** Online Information Services – Web-based services.

General Terms
Measurement; Experimentation; Human Factors; Management

INTRODUCTION
Collaborations with scientific colleagues are essential for most researchers, forming an important aspect of their career. One of the most visible acts of collaboration is co-authorship, where two or more researchers contribute to a publication. Other than co-authorship, there are several other situations that call for collaboration, e.g. assembly of a (formal) working group in a large university department or company, gathering of researchers in order to prepare a project proposal for a research grant or searching for contributors to a conference, a congress or a workshop. Studies show [8, 14, 24] that there is a need for researchers, in academic circles as well as other knowledge-intensive organizations, to find qualified collaborators. In the expanding social Web a scientist can gather useful information about his or her potential partners. Social websites in particular offer new (and more varied) information about scientists, potentially improving the researcher's decision-making. Additionally to the classical information services as the Web of Science (WoS) or Scopus a new perspective can be considered – which is the users' perspective: Where in WoS and Scopus we have the publishing scientists, in bookmarking systems we have the users of the Web 2.0, who bookmark scientific articles and assign tags to them. The users contribute to the content of the social Web and may offer potential information which might help scientists in finding appropriate collaborators.

The following approach tries to help researchers in finding the right people and to recommend scientist for potential collaboration. It is assumed that combined social information, which derives from different perspectives either of users or of scientists, leads to better recommendations: To prove this assumption three different methods and datasets are analyzed and compared to each other. Hereby social information about a researcher and his or her partners is collected not only in multi-disciplinary information services such as WoS and Scopus, based on author co-citation analysis and bibliographic coupling of authors [23, 28]. But social information is also collected in the social bookmarking service CiteULike, using collaborative filtering methods to measure researcher similarity.

Ten "target researchers" took part in the study. They evaluated the recommended similar authors in a semi-structured interview, whereby they got lists of author names and additional visualized author networks to identify potential collaborators.

The approach tries to answer two basic research questions which are at the core of how to further develop an expert recommendation system:

CSCW '13, February 23–27, 2013, San Antonio, Texas, USA.

1. Is the usage of social information from different sources more appropriate to help researchers in finding new potential collaboration partners?

2. In which way do the results based on diverse datasets and methods differ from each other and how can they be combined?

SOCIAL INFORMATION ABOUT RESEARCHERS

Similarity Measurements

It is an extremely difficult task to find serious collaborators – not just any who happen to be available but the "right" ones, proven experts with solid reputations [7]. A researcher's reputation grows with the number of publications he contributes to peer-reviewed journals, and with the frequency at which those publications are cited by others [8]. If a researcher wishes to know whether another scientist might be a good collaborator, he or she might look at the scientist's reputation and search for any relation between his or her work and the work of the scientist. He or she might ask: Who uses the same references in his or her publications like I do (bibliographic coupling measurement [19])? Or who is cited by the same authors like me (author co-citation analysis [39])? These two approaches measure the similarity between two authors. Similarity in both cases means that these scientists might be interested in the same research areas and topics due to the overlap in the citations of their works or in their usage of the same references. Similarity measurement is then used as an indicator for high collaboration potential. Of course an argument against this assumption is that researchers also search for "un-similar" collaborators who complement their own work. For example some of the ten physicist researchers, who took part in the following study, stated that they do computer simulations of physical processes. This does not mean that they only need collaborating colleagues, who also do simulations. But they might need physicists who set up experiments of their simulated processes. Another example is that a person who has established a theoretical concept might need a practitioner who can implement this concept into a real environment. The question is whether in these cases similarity measurement can be an indicator for high collaboration potential. When using bibliographic coupling and co-citation analysis as measurement method, it depends on the references and citations the authors use. If for example the theorists also cite researchers concerned with applied sciences in their published papers, these people could be recommended as potential collaborators. If they do not cite them, these people would not be retrieved in the following dataset of WoS and Scopus. Concerning the CiteULike dataset there might be a greater variety of related bookmarked papers and tags. This might lead to a greater variety of people who could be recommended to the target researcher. The next paragraph will talk about this new dataset in more detail.

In our study we gather social information based on author co-citation and bibliographic coupling in WoS and Scopus. The main aspect is in showing a target scientist new and unknown partners he or she has not recognized before. Therefore we only consider the implicit relations of author co-citation and bibliographic coupling and not the explicit relations of co-authorship and citations, for here it is certain that one author knows the other: of course one knows who one's co-authors are and, we can assume, the authors one

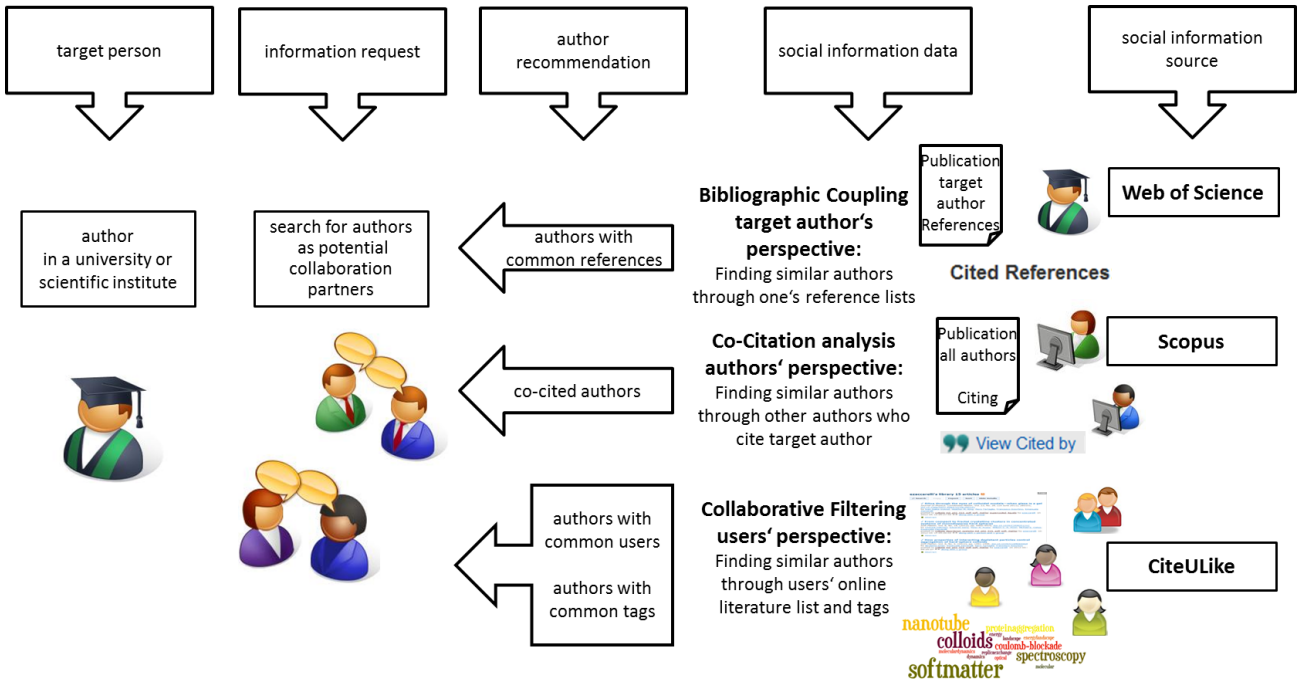

Figure 1: Concept of author recommendation based on different social information sources.

has cited. But additional social information about a researcher can be found in social bookmarking systems such as e.g. CiteULike. Here the users add social information about scientists to the services on the web. The advantage hereby is that the users' perspective is considered: Author co-citation only takes into account the perspective of a third researcher citing two other authors who might be similar. Bibliographic coupling only considers the perspectives of two authors, marked by their choice of references. Social information from web services also considers the users' perspective, as it takes into account the content they themselves have contributed to the services (Fig. 1). We thus have access to the perspectives of a large group of people. These people might be researchers themselves, but they might only be readers. That means we have the pure readers, i.e. readers who read and bookmark papers, but do not publish and hence do not cite [35]. Therefore we take into account the perspectives of people different to the publishing and citing authors in WoS and Scopus.

CiteULike is an online reference management system. The user can add article references and other Web-resource references to the system and save them as bookmarks on his or her profile. Tags, the keywords of a bookmarking system, can be assigned to references. They offer additional information and help the user sorting his or her references. The bookmark and tag lists of users are made public and are accessible by all other users, i.e. a user of CiteULike can search for a special paper or search for a special topic via the tags. The entries of the bookmarks show the resource, e.g. a scientific paper, the users who added the resource to their profile and the tags which were assigned to that resource. There is also the possibility to search for DOIs and authors of the bookmarked papers. In CiteULike users can also establish groups, e.g. open or closed research or project groups. Bookmarks added to the group's profile are then directly seen by the members of the group. As CiteULike offers a variety of functions there are different kind of users with different motivations to use this service: Users can be "common" readers, e.g. students or practitioners in a company, who use the service to manage their private references from any web-accessible device. They likewise might be scientists, who use CiteULike to manage their literature. Scientists might also be interested in distributing their own work while bookmarking their written papers. Or they are working on a project and establish a group in CiteULike to exchange potential important literature amongst their project colleagues. That means we might have those scientists in CiteULike, who might also be found in WoS and Scopus. But in CiteULike they have different motivations, i.e. here they might collect scientific papers they probably need for their work – also as papers' references, but not exclusively. So a scientist's bookmarking list might be much bigger than the reference lists of his or her published papers. Summarized we have a large group of users with different motivations; and the implicit or explicit relations between them, their bookmarks and their tags can be used to gather new social information about scientists and the relations between them.

Another aspect is that in CiteULike the researcher can get a more rapid response to his or her publications. The scientific papers can be bookmarked at the exact time when the paper is first published. When a user finds an interesting scientific article he or she can directly put it at the bookmarking list in CiteULike. This means even if a researcher's papers have not been cited yet, there might be social information about him or her in CiteULike, which leads to relations between him or her and other scientists. This aspect might particularly be important for young researchers who have only just started their academic career and have not built a scientific reputation for themselves yet. Blazek calls these "domain novice researchers", i.e. academics who enter a new domain and wish to use a collection of academic documents [4]. They face the cold start problem: Citation analysis can hardly be applied to novice researchers as long as they have little or no references and citations to their name. Furthermore, there is a time lag when measuring citations and author co-citations because an author's article will not be cited until several months after its publication, with differences in time span depending on the scientific discipline. This means that a researcher with a recently published article who might be a good collaboration partner simply will not be considered in author co-citation measurement. But if a user has bookmarked this article, the researcher can be considered when analyzing the CiteULike data.

Expert Recommender Systems

For a researcher in need of collaboration partners, a recommendation (or recommender) system could point out relevant individuals on the basis of various characteristics. Nowadays, recommender systems use different methods and algorithms for different items, e.g. products, movies, music, articles, etc., the goal being personalized recommendation [3] of items unknown, yet relevant, to the target user. The first question is where to find the best resources for a user and how to rank them according to their relevance [12]. One method is collaborative filtering (CF), which not only considers the ratings of a unique user, but also those of other users [a.o. 16, 30, 34]. One advantage of CF compared to the content-based method is that recommendations rely not only on the item's content, which may be insufficient for quality indication, but also on the evaluation of other users. When using CF to recommend potential collaborators to a target researcher, taking into account the users' perspective should yield new and more appropriate results. It is the principle of the Amazon-style recommendation. If, for instance, many users have added works by two certain authors to their online literature list in a social bookmarking system, this would be a further hint that these authors share similar research interests – assuming, of course, that the users are interested in a

specific topic and also bookmark the relevant literature. This social relation is similar to author co-citation, but assigned to a social bookmarking service, with the difference that we don't necessarily have scientific authors, but users of the social web. The difference lies in the new dataset: Not only do we now have the opinion of more people, but this new user perspective may also expand a researcher's known social network and uncover new relations between researchers.

Recommender systems work by assigning user ratings to the items, a method called user-item response [12]. In our approach we use CiteULike, a social bookmarking and tagging system, where users are now also able to rate the bookmarked scientific articles. But instead of these ratings we consider the unary user-item response, which means that while a user has not rated an item, his purchase of or access to the item is interpreted as a positive response – and this response can be used for recommendations [15, 25].

There are several studies that investigate expert recommendation, e.g. for commercial enterprises [6, 32, 33]. Petry et al., for instance, have developed the expert recommendation system ICARE, meant to recommend experts within an organization [32]. In this system the primary spotlight is not directed onto an author's publications and citations, but rather on their organizational level, availability and reputation, among other aspects. Following a field study and interviews with employees, Reichling and Wulf explore the options of a recommender system to support their knowledge management [33]. In this system experts are defined via their collection of written documents, which have been analyzed automatically. The authors also used a post-integrated user profile with information about each individual's background and job description. The use of user profiles in bookmarking services might also improve the effectiveness of user recommendation.

In addition to user recommendation for commercial enterprises, several other approaches concentrate on Web 2.0 users and academics [10, 11]. Au Yeung et al. discuss the non-academic bookmarking system Del.icio.us and define an expert user as someone, who has deposited high-quality documents in their bookmark collection (many users who have high levels of expertise fulfill this criterion) and who tends to recognize useful documents well in advance of others (as seen in the timestamps on users' bookmarks) [1]. In contrast to the following approach, the "high-quality documents" in this experiment are the publications of the researcher to whom collaboration partners are meant to be recommended. Hence, it is vital for the purposes of recommendation that users bookmark at least one of the author's publications. Heck and Peters propose using social bookmarking systems for scientific literature such as BibSonomy, CiteULike and Connotea to recommend researchers unknown to the user, but who share the same interests and would thus be suitable partners for building a community of practice [14]. Users are recommended to each other when they have either bookmarks or tags in common. Ben Jabeur et al. [2] use social clues such as the connectivity of researchers and opportunities to meet in person, e.g. at scientific conferences, to improve the performance of the recommendation system. Nocera and Ursino focus on "social folksonomy", i.e. using information about user friendships and semantic information of tags [29] for their recommendations. What we consider in our approach are not the explicit relations between two researchers like in [2], but the indirect relations because we assume that they would provide more unknown potential collaboration partners for recommendation. Furthermore in our case a researcher does not need to be active in a social service, i.e. he or she does not have to bookmark his or her articles, but the users do this job. Nevertheless if a researcher bookmarked his or her own publications, this might increase the chance of other users also bookmarking these articles when they search for literature in the bookmarking system. This again would influence the relations between the researcher and other authors and might lead to new recommendations.

Another important aspect for recommender systems is their evaluation. Recommender systems should not only prove accurate and efficient, but must also detect the users' respective needs in order to be of use to them [17]. Several studies incorporate user evaluation in their investigation of the evaluation of model-based recommender systems [20, 26]. McNee et al. show the pitfalls of recommender systems in order to foster user acceptance and promote further usage of recommender systems as knowledge management tools [27]. The following example of a recommender system is also evaluated by the target researchers.

SOCIAL INFORMATION ABOUT RESEARCHERS

Collaborative Filtering in CiteULike

Unlike bookmarking systems such as Del.icio.us, CiteULike focuses on the management of academic literature. The basis for social recommendation is the service's folksonomy structure. A folksonomy [25, 31] is defined as a tuple F: = (U, T, R, Y), where U, T and R are finite sets with the elements of 'user name', 'tag' and 'resource' and Y is a ternary relation between them: $Y \subseteq U \times T \times R$ with the elements being called 'tag actions' or 'assignments'. To use this information for recommending authors to each other, we expand the folksonomy to FE:=(U, T, R, A, Y), where A is added as the finite set with the element 'authors' and $Y \subseteq U \times T \times R \times A$ is their relation.

In our experimental comparison, we want to cluster scientific authors with similar research interests. We are not interested in the networks and relations of the CiteULike users themselves but only in their bookmarks, i.e. the bookmarked publications of our target scientists, and the tags assigned to those bookmarks. Therefore we have two

options for setting our database for author similarity measurement:

1. Searching for all users $u \in U$ who have at least one article by the target author a in their CiteULike bookmark list.

2. Searching for all resources $r \in R$ that have tags $t \in T$ in common with one bookmarked article $s \in R$ by our target author $a \in A$.

The disadvantage of the first method, for us, lies in the small number of users. Relying only on the users may not be enough to identify similarity [21]. The disadvantage of the second method is that users who have an article of target author a in their bookmark list, but have not tagged it, might get lost and not be in the dataset. But as many users tagged the articles of our target authors, we decided to use the second method (except of two cases: see experimental results): We searched for all resources which have been assigned the same tags as the resources of our target researcher. Resources (here: scientific papers) can be deemed similar if they have been assigned some shared tags. From here, we assume that the authors of these documents are also similar. Tags point to topical relations, i.e. authors connected via such relations regarding their research fields can be potential collaboration partners. Additionally, the more tags are shared by two documents, the more similar they are. In some cases, our target authors' articles were labeled with very general tags such as "nanotube" and "spectroscopy", so we decided to determine a minimum of two unique tags that a document must have in common with a target author's document.

To measure similarity we use the cosine coefficient. In our dataset we measure author similarity in two different ways: (A) Based on shared tags assigned to the resources of target author a and author b; (B) Based on shared users who have bookmarked the resources. The similarity between authors a and b is measured:

$$A) sim(a,b) := \frac{|T_a \cap T_b|}{\sqrt{|T_a| * |T_b|}} \qquad B) \mathrm{sim}(a,b) := \frac{|U_a \cap U_b|}{\sqrt{|U_a| * |U_b|}} \qquad \text{Eq.(1)}$$

where T_a, respectively T_b, is the set of tags assigned to the scientific papers of target author a, respectively author b, and U_a, respectively U_b, is the amount of users having bookmarked scientific papers of a, respectively b.

Author Co-Citation in Scopus

Co-citations [22, 36, 37, 39] are undirected weighted linkages between two scientific papers, calculated via their fraction of co-citations. We then aggregate the data from the document level to the author level.

In Author Co-Citation (ACC), two authors a and b are linked if they are cited in the same documents. We cannot use WoS to mine author-co-citation data because only the first author of any cited document is listed in the reference section of its bibliographic entries, whereas we require a complete list of all authors [41]. Therefore we will mine this data from Scopus, for here we find more than one author of the cited literature. We perform an inclusive all-author co-citation, in which two authors are considered co-cited when another author cites a paper they co-authored [40].

In Scopus we search for all documents that cite at least one of the target author's articles. We analyze the references of these documents to find authors who are co-cited with our target researcher. For similarity measurement using the cosine coefficient we divide the number of documents co-citing the two authors by the product of both authors' citations (see Eq. (1)). We did not consider all authors who were co-cited with a target researcher. The main reason is the problem of ambiguous author names in the data (see paragraph about data limitations). To get a good quality of our dataset we searched for the co-cited authors and for their citations manually. Thus we limited the number of authors who are considered for similarity measurement: We ranked authors according to the total number of co-citations with our target researcher. Then we took those authors with a minimum number of x co-citations, where x determined the cutting point of the ranking. This number of co-citations may vary for each target scientist ($x >= 3$). By applying this procedure different numbers of authors are considered for similarity measurement for each of the ten target researchers (minimum 34, maximum 69 authors).

Bibliographic Coupling in WoS

BC is also an undirected weighted linkage between two scientific papers, calculated via their fraction of shared references. BC of authors means that two authors a and b are linked if they cite the same authors in their papers. We use WoS to mine data about bibliographic coupling, since this service allows searches for "related records", where relations are calculated via the number of references a certain document has in common with the source article [38].

Our assumption is: Two authors with one document each that share a high number of identical references are more similar than two authors with a large amount of shared references across many documents – the number of shared references per document being the vital quantity. For example: Let author a have 6 references in common with authors b and c. These 6 shared references are found in two unique documents by author a and author b respectively, but for author c they are distributed across 6 individual documents. In this scenario authors a and b are more similar than authors a and c, because the reference lists of *a's* and *b's* documents are more similar.

Therefore in WoS we searched for all related documents that share at least n references with any of the publications by target author a, where n may vary for each target scientist. We took the authors of these documents for the BC measurements. Authors of multiple documents in the

dataset were summarized. After this step we searched for the number of authors with common references and their number of references in WoS manually to improve the quality of the dataset. Therefore we limited the number of authors who are considered for the similarity measurements: We ranked authors according to the total number of common references with our target researcher. Then we took those authors with a minimum number of *y* common references, where *y* determined the cutting point of the ranking. This number of common references may vary for each target scientist ($y >= 4$). By applying this procedure different numbers of authors are considered for similarity measurement for each of the ten target researchers (minimum 22, maximum 53 authors). We measure the similarity between these authors and our target scientist via the cosine coefficient and divide the amount of references shared by both authors by the product of the references of both authors (see Eq. (1)). Regarding the results of research literature, the best performance in terms of representing research activities is achieved by both methods, BC and ACC, in combination [5, 13].

Construction of Similarity Networks and Graphs

Applying the proposed three mined datasets and four similarity approaches, we are able to assemble four different networks of potential similar authors. One network is based on BC in WoS (BC-network), one is based on ACC in Scopus (ACC-network), one on CF of shared users in CiteULike (CULU-network), and the final network is based on CF of shared tags in CiteULike (CULT-network). We can now analyze those authors who, according to the cosine coefficient, are most similar to our target author, and evaluate the results. Also on the basis of the mined datasets we can measure the similarity between all authors of a network. These results are shown to the evaluators in visualizations, which we call graphs (example Fig. 3). Therefore, a visualized graph exists for each network and will likewise be evaluated. We used the Gephi software (www.gephi.org/) for network visualization. The size of the nodes (=author names) depends on number of relations of an author, the edges are sized according to the cosine weight. Note that the CiteULike graphs are much larger, since we cut off the list of related authors in WoS and Scopus. To get a clear graph arrangement for a better evaluation, we set thresholds based on the cosine coefficient when needed. Additionally, we left out author pairs with a similarity of 1 if they had only one user or tag (in the CiteULike dataset) in common, as this would have distorted the results.

Limitations of Data Collection

A good recommendation system highly depends upon the source dataset. Various problems arise while filtering information in the three information services. We will briefly discuss these problems: One of the main problems is the occurrence of ambiguous author names and different name spellings. We tried to identify unique authors mainly through co-authors and research field areas (e.g. subject areas in WoS). At best we could identify an author through his unique author-ID like in Scopus or at the website researcher-ID (www.researcherid.com: linked with an author's data in WoS and searchable in the service since 2012). Nevertheless author details in Scopus are not always correct. In CiteULike besides the diverse spellings of an author name, users might make mistakes and misspell the names and article titles. An all three sets we tried to detect the mistakes and corrected the author lists manually.

Furthermore in Scopus, we detected differences in the metadata: One and the same article may appear in several different ways, i.e. title and authors may be listed completely in one reference list but incompletely in the references of another article. In our case, several co-authors in the dataset went unmentioned and could not be considered for author co-citation analysis.

EVALUATION

Target Scientists and Interview Structure

We cooperated with the *Forschungszentrum Jülich, Germany*, and ask physicists to provide us their correct and complete publications lists. Ten researchers answered to be available for the interview. For those ten physicists we built individual networks and graphs. There were two phases of data collection and evaluation: One in May 2011 (evaluation by 6 authors) and one in May 2012. We limited the source for the dataset modeling to the authors' publications between 2006-2011, respectively in the second phase between 2006-2012, in order to make recommendations based on the physicists' actual research interests. While modeling the datasets, we found that one of the ten authors didn't have any users who bookmarked his articles in CiteULike. Some articles were found but they had been adjusted to the system by the CiteULike operators themselves, so the CiteULike networks couldn't be modeled for this scientist. Another researcher's articles were bookmarked, but not tagged. In this case we searched for all users who had bookmarked his articles (instead of all resources that have tags in common with his articles).

The main question the evaluators should answer was, if a recommended author is relevant for the evaluator's current research and if he or she could be a good collaborator. Another aspect, as we stated before, is that we want to recommend relevant, unknown potential partners the target researcher has not recognized before. The difficulty in this case is that a person cannot make a statement about another person's relevance if this person is unknown to him or her. The evaluator should at least know one or more scientific papers of the author he or she should evaluate. To evaluate the relevance of the recommended unknown people, one could have taken recent research papers written by these scientists. On the basis of the papers an interviewee could have made his or her opinion about the relevance of the authors of these papers. Due to lack of time during the

interviews and the determined structure of those we did not consider this option, but will preserve it for further research.

For the interview we took the top ten most similar authors of all four measurements according to the cosine coefficient and listed them in alphabetical order (co-authors being eliminated). The evaluators should distinguish between two types of persons: First authors who were new and unknown to the evaluator (here the physicist was not able to make a statement about the relevance if this author and his or her appropriateness as a collaborator). Second authors the evaluator knows because he or she reads the author's papers or he or she is even personally acquainted to the author. Here the target researcher could make a statement about the author's relevance as a collaborator, although in some cases the evaluator stated that he or she doesn't know the author well enough and couldn't make a clear statement about the relevance. The target authors had to rank the known authors according to their relevance (rating from not important at all (1) to very important (10)). Note that only the known authors could be ranked. Additionally a target scientist could name authors who are not on the list, but are important for his or her current research.

The target scientists should also evaluate the graphs, the visualized networks with the relations between all authors based on the cosine coefficient. The graphs should be examined according to the right distribution of the authors and their relations to each other. Here a main aspect was the target author's own relations and the placement within the graphs' author network. Furthermore the interviewees should rate the graphs (rating from not important at all (1) to very important (10)) concerning the question whether a graph is helpful to find relevant partners, e.g. to organize a conference or workshop or for collaboration activities.

Relevance of Recommended Authors

For the recommended authors we have two groups, which could be distinguished: unknown authors and known authors. In almost all cases (except three) the number of unknown authors is higher than the number of known authors. The interviewees knew 211 recommended authors 166 recommended authors were unknown to the target researchers. It seems that senior scientists already have a huge collaboration network, as they also knew many authors personally. There were two reasons why the known authors would not be good collaborators: Either there was less topical overlap between the researcher's and the recommended author's work or the recommended author was rather a competitor than a collaborator.

On the first sight the relatively high number of unknown authors seems to be a good result as we want to recommend new potential collaborators to our target scientists. On the other hand the researchers stated that they are interested in new collaboration partners, but would prefer people of whom they at least know their scientific work. It seems that

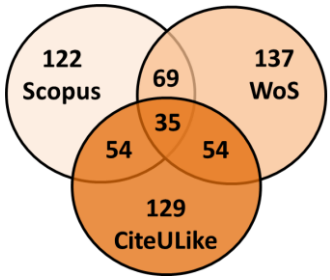

Figure 2: Distribution of important authors for all scientists.

personal familiarity has a great influence on the decision whether a person is relevant to be a collaboration partner or not. This makes it difficult for a recommender system to choose the right people. Totally unknown people wouldn't likely be good collaborators, as the scientists couldn't evaluate their reputation and appropriateness of their work. Therefore a network only showing unknown authors would not be helpful for the scientists. To overcome this difficulty, links to a recommended author's published papers or web profile could help. With additional information about the author the target scientist could estimate his or her relevance in a better way. Another hint of good collaborators, which was claimed by the interviewees, is the relation of unknown people to relevant people who are known by the target author. Thus we decided to have a closer look at the identified relevant and important authors, who would be good collaborators.

To have a closer look at the important authors, we used the ratings the target researchers gave to the known authors: We defined important authors as authors with at least a rating of 5 who were deemed important by an interviewee, as well as all important authors added by the interviewee which were not on any network's Top 10 list. In general, our target authors named 18 till 55 people they regarded as important for their recent scientific work. Note that in some cases the target author changed his or her research field. Thus some known recommended authors were relevant collaborators in the past, but not appropriate for the current research. The general distribution of important authors in the three services is shown in Figure 2. It can be recognized that in all three datasets almost the same number of important authors is found, but that the overlapping is less, i.e. diverse important scientists are found using the diverse measurements in WoS, Scopus and CiteULike. Furthermore it is interesting to take a look at the important authors who were only found in the CiteULike networks: For example, 7 of the 35 important authors for target author *a2* are only found in the CiteULike-network, respectively 6 of the 29 important authors of *a3*. This approves our assumption that with the use of more diverse social information more relevant authors can be found. Thus a target researcher can get better recommendations.

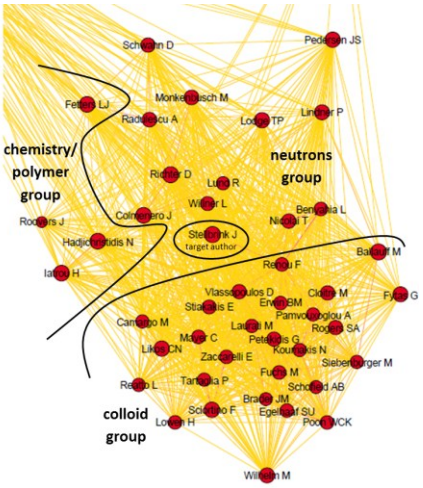

Figure 3: Extract of BC-graph with researcher groups detected by target author (in circle).

Relevance of Visualized Graphs

We assume from the statements of the evaluators that a network which shows known relevant authors and unknown people connected to the relevant ones would help an author to find new collaborators. To prove this we let the target scientists evaluate the graphs. The interviewees had to evaluate the graphs (from 1 till 10) concerning the question whether a graph is helpful to find relevant partners. The average graph relevance is:

- BC-graph: 7.95
- ACC-graph: 6.35
- CULT-graph: 4.81
- CULU-graph: 3.75

Six of the target authors claimed the BC-graph to be the best one. This ranking is not surprising since the interviewees stated to find good collaborators under the recommended, but already known authors. And with the BC method in WoS the most important authors were found. On the other hand the CULT-graph was claimed to be quite good. The authors suggested combining the BC- and the CULT-graph in order to be yielded all important authors and relevant research communities. The reason is that the BC-graph was said to show the main relevant authors, while the CULT-graph was said to reveal relations between authors of which the target scientists were not aware of before.

From the statements of the researchers we could detect two main aspects which influence the relevance of a graph. One aspect is the need of author recommendation: For the organization of e.g. a conference or workshop the scientists preferred rather bigger graphs with more unknown people. That means the CiteULike graphs could help them finding unrecognized scientists to be invited to a conference. In the BC- and ACC-graphs most interviewees stated that the authors in these two graphs were too obvious to be similar and said they were interested in bigger graphs featuring more potential collaborators. The graphs which only show relevant but already known authors might not be appropriate for recommendation, if there is a need for more unknown researchers to get new collaborators. On the other hand some interviewees said that a small network would be sufficient for them because they preferred to have a clear structure. This was the case when the researcher either does a lot of work on his or her own and is not in need of many new collaborators, or when the researcher has already built up a sufficient network and prefers the known scientists. Thus in a running recommendation system it would be good to enable users to choose between a smaller or bigger network (visualized graph). So they could decide either to stick with a smaller group of more known people or to widen their network to possibly find new collaborators. In any case the advantage of a visualized graph is that it shows the relations between the authors, which help the target scientist to recognize the relevance of a person. A simple list of recommended authors could not fulfill this aspect.

Looking at the graphs, all target authors remembered important colleagues that they had not thought of at first but whom they found very helpful. To give a clearer statement about new potential relevant authors unknown to the target researchers, the interviewees needed to have to look at their publications. But it may be assumed that if an unknown person is shown to be clearly connected to a known relevant research group, he or she probably does similar (and relevant) work. As the interviewees stated that the distribution of the researchers is shown correctly, it is likely, albeit not explicitly proven, that any unknown scientist would also be allocated correctly within the graph.

The second aspect which influences the graph rankings was the structure of the graphs themselves. At the beginning we assumed the novice researchers to suffer from the cold start problem. In fact it is a problem also by senior researchers. The coverage of the researchers' articles differs in Wos, Scopus and CiteULike, i.e. the data for the construction of the graphs also differs. It is surprising that CiteULike seems to have more articles from novice researchers and also more recent articles. To have a good structured graph the target authors' articles need to have users who bookmarked them and also assigned appropriate tags to them. If that's not the case, the huge CiteULike graphs get very unstructured and are thus not helpful for detecting collaborators and relationships among authors. If the graph had a clear structure which showed author communities and separated small groups, the researcher claimed the graph helpful in detecting author "hubs" and in finding new known and unknown collaborators. Then the target scientists could identify the important known authors and look at their relations to find unknown persons. If the graph was not well structured, which was the case if e.g. only a few tags were assigned to the target authors' articles in CiteULike, the

identifications of important authors and groups was difficult. Then the interviewees claimed the graph not to be helpful in finding collaborators. Further categorizations, e.g. via tags or author keywords, might help to classify the scientists' work and avoid unclear distributions in a graph. During the graph evaluation the researchers already arranged author groups within the graph with the help of keywords (see Fig. 3). The tags assigned to the articles of the target authors were quite appropriate and could also be helpful to detect such author groups. Therefore it would be interesting to further analyze the quality of the users' tags.

To summarize the findings from the graph evaluation we could say that the BC and ACC graphs showed a quite good amount of potential collaborators which the target researchers found helpful. They stated that these graphs cover the core of their research field and showed many known people. On the other hand the CiteULike graphs, which were quite bigger, gave a wider overview of related research fields and their authors. CULU and CULT were thus more helpful if a researcher looks for many new collaborators, e.g. for a conference, or if the researcher has different research fields which should be covered in a graph. Two target authors stated that they are working in multidisciplinary research fields and none of the BC and ACC graphs covered all of the fields. Those scientists preferred the CiteULike graphs.

Discussing the Combination of Methods

Having discussed four different methods to help researchers find potential collaboration partners, there is the question if these methods should be combined in a recommendation system? One way is just to rank the recommended authors according to their similarity to the target author (cosine coefficient). For authors who are found in more than one network, one could sum up the cosine coefficients. But the difficulty is the great difference of the cosine values. In general, it can be seen that the cosine coefficient for BC is very low compared to that for ACC as well as similarity measurements in CiteULike. The highest cosine value for BC is 0.68, but in general the interval for a researcher is between 0.3 and 0.01. This is because some authors have a lot of references, which minimizes similarity. Additionally, similarity is comparatively high for measurements in CiteULike because the number of users and assigned tags related to the target authors' publications was relatively low. Here the maximum cosine value is 1. If we want to combine the results of the clusters, the summation and ranking of the cosine values would not be appropriate. For unique authors who were found in several networks we tried to normalize the cosine values (see example in table 1). The idea was that these authors should be ranked higher because it is assumed that an important author found in more than one dataset is highly relevant. That means e.g. if the target author's (BC), the other authors' (ACC) and the users' perspective (CF) brings together two authors through implicit relations, this indicates a high similarity between both. So the more implicit relations two authors have the more similar they are.

Important authors	norm Cos	rank mod Cos	cos ACC	rank ACC	cos BC	rank BC	cos CULU	rank CULU	cos CULT	rank CULT
Wales DJ	3.20	1	0.39	2	0.37	1	0.46	36	0.59	1
Wenzel W	1.92	3			0.02	47	0.87	2	0.51	3
Klenin K	1.86	4					0.87	1	0.51	2
Carr JM	1.83	5			0.22	2	0.71	3	0.24	89
Stock G	1.62	7	0.09	26	0.18	3	0.33	61	0.36	17
Derreumaux P	1.51	8	0.15	9	0.15	4	0.29	67	0.30	28
Klimov DK	1.48	9			0.09	30	0.50	11	0.39	7
Johnston RL	1.32	10					0.71	4	0.30	38
Nguyen PH	1.31	14	0.09	22			0.38	42	0.42	5
Miller MA	1.25	16					0.71	5	0.25	66
Caflisch A	1.15	25	0.05	59	0.12	9	0.22	124	0.29	44
Pande VS	1.12	30			0.13	7	0.32	64	0.24	93
Scheraga HA	1.00	44			0.09	32	0.26	99	0.27	47
Mu Y	0.84	62					0.35	47	0.25	67
Karplus M	0.82	64			0.10	21	0.17	191	0.21	143
Brooks CL	0.77	72			0.10	17	0.11	249	0.22	133

Table 1: Comparison: normalized cosine and cosine values and ranks for 16 important authors of one target scientist.

In the example the important authors got quite good ranks with the normalized and summated cosine values. On the other hand we showed that there are many important authors who are only found in one network (Fig. 2). The weakness of combining the datasets to only one big network is that these relevant people would not be found under the top ranks. Especially if a user is shown a rank list, he or she will probably not pay attention to the lower ranked authors. In a visualized graph all authors would appear and can be detected by the target researchers. However one challenge in a visualized graph, which shows results from all networks, will be the changing of the relations: That means that the relationships between authors and author groups will change. The target author might not be able to detect relevant persons and groups anymore. If the cosine values change, e.g. if an author-author-relation appears in more than one network and the different similarity cosine values are combined, the visualization of the network may change enormously. This will influence the perception of the target author concerning relevant persons and researcher groups. Therefore it should be evaluated if the relations of such a combined graph are still correct. Another great issue which should be considered when combining the networks is the problem of author ambiguity, which would be tough to handle when using different services with diverse spelling preferences. Thus it is arguable whether the combination of the methods, i.e. the cosine values, leads to any new advantages to the target author.

CONCLUSION

Regarding the need of a researcher to collaborate with his or her colleagues, social information can be used to build networks of researchers and to recommend similar people to each other. It is important for a service to recommend the right people in order to satisfy its users and be of advantage to them. For recommendations, the reputation of the potential partner is very important, hence citations and references should be considered. Our approach combined

scientometric measurements with the CF-method and used data across three information services. We could show that user-generated social information in a bookmarking service complements author co-citation and bibliographic coupling measurements. All three measurements showed different results, i.e. different relevant authors were found and the number of relevant authors increased with the use of the three datasets. The graphs gave deeper insights into the usefulness of author recommendation: In the graphs the target researchers detected relevant collaborators whom they didn't recognize before. This aspect was found very helpful in finding new partners. On the other hand, the interviewees preferred networks with known relevant authors, who reveal other relations to unknown and potential knew collaborators. The relevant known authors function as orientation to find new relevant collaborators. A visualized network or graph is thus more helpful than a recommended author list because it doesn't only reveal author names, but also shows the relations between the authors. Nevertheless it also depends on the target researchers needs and preferences, which methods and graphs lead to good author recommendations.

In the following research we want to test our methods against other similarity measurements and methods to improve author recommendation. Besides other combinations of social information data can be tested. Further investigations testing the best algorithm for similarity measurement are required. The relations between user-based and tag-based similarity in a bookmarking system should also be considered, e.g. via a graph-based approach such as FolkRank [18] or user expertise analysis [1]. Aspects such as accuracy and efficiency should be tested in an operating recommendation system. Apart from technical aspects, the target users – for example, researchers – should be involved in the system's evaluation. If they don't see its benefits and don't trust its recommendations, the service won't be of any use.

ACKNOWLEDGMENTS

I would like to thank Wolfgang G. Stock, Isabella Peters and Oliver Hanraths for their contributions to the paper, and the physicists for their evaluation. The project was financed by a grant of the *Strategische Forschungsfonds* of the Heinrich-Heine-University Düsseldorf.

REFERENCES

1. Au Yeung, C. M., Noll, M., Gibbins, N., Meinel, C., and Shadbolt, N. On measuring expertise in collaborative tagging systems. In *Proc. Web Science Conference: Society On-Line.* (2009). Retrieved from http://journal.webscience.org/109.
2. Ben Jabeur, L., Tamine, L., and Boughanem, M. A social model for literature access: towards a weighted social network of authors. In *Proc. RIAO 2010,* ACM Press (2010), 32-39.
3. Berkovsky, S., Kuflik, T., and Ricci F. Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction, 18,* 3 (2007), 245-286.
4. Blazek, R. *Author-Statement Citation Analysis Applied as a Recommender System to Support Non-Domain-Expert Academic Research* (Doctoral Dissertation). Fort Lauderdale, FL: Nova Southeastern University (2007).
5. Boyack, K. W., and Klavans, R. Co-citation analysis, bibliographic coupling, and direct citation. Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61,* 12 (2010), 2389-2404.
6. Cai, X., Bain, M., Krzywicki, A., Wobcke, W., Kim, Y. S., Compton, P., and Mahidadia, A. Collaborative filtering for people to people recommendation in social networks. In J. Li (Ed.), *LNCS: Vol. 6464. Advances in Artificial Intelligence*, Berlin, Germany: Springer-Verlag (2011), 476-485.
7. Cronin, B. *The Citation Process. The role and significance of citations in scientific communication*. London, UK: Taylor Graham (1984).
8. Cronin, B., Shaw, D., & La Barre, K. A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology, 54,* 9 (2003), 855-871.
9. Cruz, C. C. P., Motta, C. L. R., Santoro, F. M., and Elia, M. Applying reputation mechanisms in communities of practice. A case study. *Journal of Universal Computer Science, 15,* 9 (2009), 1886-1906.
10. De Meo, P., Nocera, A., Terracina, G., and Ursino, D. Recommendation of similar users, resources and social networks in a social internetworking scenario. *Information Sciences, 181* (2011), 1285-1305.
11. Deng, H., King, I., and Lyu, M.R. Formal Models for Expert Finding on DBLP Bibliography Data. In *Proc. ICDM 2008,* ACM Press (2008), 163-172.
12. Desrosiers, C., and Karypis, G. A comprehensive survey of neighborhood-based recommendation methods. In F. Ricci, L. Rokach, B. Shapira, & P.B. Kantor (Eds.), *Recommender Systems Handbook*. New York, NY: Springer-Verlag (2011), 107-144.
13. Gmur, M. Co-citation analysis and the search for invisible colleges. A methodological evaluation. *Scientometrics, 57,* 1 (2003), 27-57.
14. Heck, T., and Peters, I. Expert recommender systems: Establishing Communities of Practice based on social bookmarking systems. *Proc. I-Know 2010,* (2010), 458-464.

15. Heck, T., Peters, I., and Stock, W.G. Testing collaborative filtering against co-citation analysis and bibliographic coupling for academic author recommendation. In *Proc. RecSys'11 Workshop on Recommender Systems and the Social Web*. (2011). Retrieved from http://www.dcs.warwick.ac.uk/~ssanand/RSWeb11/index.htm.
16. Herlocker, J. L., Konstan, J. A., Borchers, A., and Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In *Proc. SIGIR 1999*, ACM Press (1999), 230-237.
17. Herlocker, J. L., Konstan, J. A., Terveen L. G., and Riedl, J. T. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22,* 1 (2004), 5-53.
18. Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G. Information retrieval in folksonomies: Search and ranking. In Y. Sure & J. Domingue (Eds.), *LNCS: Vol. 4011. The Semantic Web: Research and Applications,* Heidelberg, Germany: Springer-Verlag (2006), 411-426.
19. Kessler, M. M. Bibliographic coupling between scientific papers. *American Documentation, 14* (1963), 10-25.
20. Krohn-Grimberghe, A., Nanopoulos, A., and Schmidt-Thieme, L. A novel multidimensional framework for evaluating recommender systems. In *Proc. RecSys 2010 Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces*, (2010). Retrieved from http://ucersti.ieis.tue.nl/2010/program.html.
21. Lee, D. H., and Brusilovky, P. Social networks and interest similarity. The case of CiteULike. In *Proc. HT 2010,* ACM Press (2010), 151-155.
22. Leydesdorff, L. Similarity measures, author cocitation analysis, and information theory. *Journal of the American Society for Information Science and Technology, 56,* 7 (2005), 769-772.
23. Li, J., Burnham, J. F., Lemley, T., and Britton, R. M. Citation analysis. Comparison of Web of Science, Scopus, SciFinder, and Google Scholar. *Journal of Electronic Resources in Medical Libraries, 7,* 3 (2010), 196-217.
24. Luukkonen, T., Persson, O., and Sivertsen, G. Understanding Patterns of International Scientific Collaboration. *Science, Technology, & Human Values, 17,* 1 (1992), 101-126.
25. Marinho, L. B., Nanopoulos, A., Schmidt-Thieme, L., Jäschke, R., Hotho, A., Stumme, G., and Symeonidis, P. Social tagging recommenders systems. In F. Ricci, L. Rokach, B. Shapira & P.B. Kantor (Eds.), *Recommender Systems Handbook*. New York ,NY: Springer-Verlag (2011), 615-644.
26. McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., and Riedl, J. On the recommending of citations for research papers. In *Proc. CSCW 2002*, ACM Press (2002). 116-125.
27. McNee, S. M., Kapoor, N., and Konstan, J.A. Don't look stupid. Avoiding pitfalls when recommending research papers. In *Proc. CSCW 2006,* ACM Press (2006). 171-180.
28. Meho, L. I., and Sugimoto, C. R. Assessing the scholarly impact of information studies. A tale of two citation databases – Scopus and Web of Science. *Journal of the American Society for Information Science and Technology, 60,* 12 (2009), 2499-2508.
29. Nocera, A., and Ursino, D. An approach to provide a user with recommendations of similar users and potentially interesting resources. *Knowledge-Based Systems, 24,* 8 (2011), 1277–1296.
30. Parra, D., and Brusilovsky, P. Collaborative filtering for social tagging systems. An experiment with CiteULike. In *Proc. RecSys 2009,* ACM Press (2009), 237-240.
31. Peters, I. *Folksonomies. Indexing and Retrieval in Web 2.0*. Berlin, Germany: De Gruyter Saur (2009).
32. Petry, H., Tedesco, P., Vieira, V., and Salgado, A. C. ICARE. A context-sensitive expert recommendation system. In *Proc. Workshop on Recommender Systems on European Conf. on Artificial Intelligence*, (2008), 53-58.
33. Reichling, T., and Wulf, V. Expert recommender systems in practice. Evaluating semi-automatic profile generation. In *Proc. CHI 2009*, ACM Press (2009), 59-68.
34. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. Grouplens: An open architecture for collaborative filtering of netnews. In *Proc. CSCW 1994,* ACM Press (1994), 175-186.
35. Rowlands, I., Nicholas, D. The missing link: journal usage metrics. In *Proc. Aslib, 59, 3* (2007), 222-228.
36. Schneider, J. W., and Borlund, P. Matrix Comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology, 58,* 11 (2007), 1586-1595.
37. Schneider, J. W., and Borlund, P. Matrix Comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics. *Journal of the American Society for Information Science and Technology, 58,* 11 (2007), 1596-1609.
38. Stock, W. G. Web of Science. Ein Netz wissenschaftlicher Informationen – gesponnen aus Fußnoten [Web of Science. A web of scientific information – cocooned from footnotes]. *Password, no. 7+8* (1999), 21-25.

39. White, H. D., and Griffith, B. C. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, *32,* 3 (1981), 163–171.

40. Zhao, D., and Strotmann, A. All-author vs. first author co-citation analysis of the Information Science field using Scopus. *Proceedings of the 70th Annual Meeting of the American Society for Information Science and Technology, 44,* 1 (2007), 1-12.

41. Zhao, D., and Strotmann, A. Counting first, last, or all authors in citation analysis. Collaborative stem cell research field. *Journal of the American Society for Information Science and Technology, 62,* 4 (2011), 654-676.