You Never Walk Alone: Recommending Academic Events Based on Social Network Analysis

Ralf Klamma, Pham Manh Cuong, and Yiwei Cao

Databases & Information Systems
RWTH Aachen University, Ahornstr. 55, D-52056 Aachen, Germany
{klamma,pham,cao}@dbis.rwth-aachen.de

Abstract. Combining Social Network Analysis and recommender systems is a challenging research field. In scientific communities, recommender systems have been applied to provide useful tools for papers, books as well as expert finding. However, academic events (conferences, workshops, international symposiums etc.) are an important driven forces to move forwards cooperation among research communities. We realize a SNA based approach for academic events recommendation problem. Scientific communities analysis and visualization are performed to provide an insight into the communities of event series. A prototype is implemented based on the data from DBLP and EventSeer.net, and the result is observed in order to prove the approach.

Keywords: Recommender systems, Social Network Analysis, community analysis, community of practice, information visualization.

1 Introduction

Academic events play an important role as the major publication and dissemination outlet in scientific communities. In computer science, the number of academic events has increased dramatically in recent years, which is evident in data from DBWorld[1]collected by [6] and data from DBLP and EventSeer.net (see Figure 1). It is challenging, especially for young researchers to find suitable events for submitting papers to and to join in some research communities. There is also the need to identify the research community of a particular researcher.

Until now, tools and methodologies developed for academic events management and documentation still have problems. Event management systems consider event managing process from event announcement, paper submission, paper review to paper acceptance notification. Digital libraries like ACM[2], DBLP[3] or CiteSeer[4] mainly focus on research publications providing tools for papers

[1] http://www.cs.wisc.edu/dbworld/
[2] http://portal.acm.org/dl.cfm
[3] http://www.informatik.uni-trier.de/~ley/db/
[4] http://citeseerx.ist.psu.edu/

J. Zhou (Ed.): Complex 2009, Part I, LNICST 4, pp. 657–670, 2009.

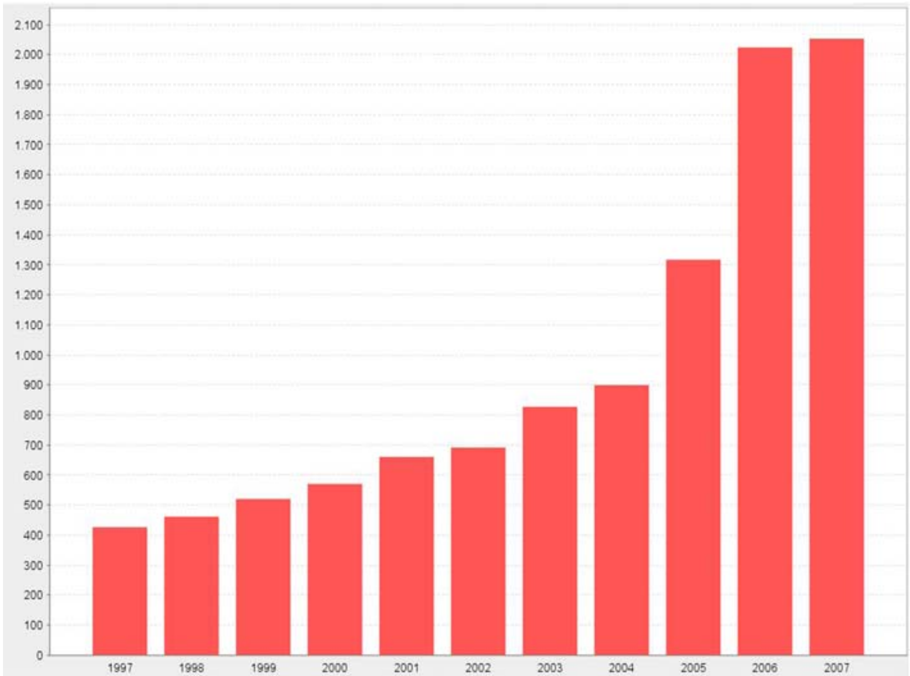

Fig. 1. Number of events in DBLP (by distinct proceedings)

searching. Some other systems like EventSeer.net[5] make a step forward to academic event and community analysis. None of the above mentioned systems provides recommendation tool to help researchers in event finding.

To overcome the aforementioned problem, a model for academic events is required. Event and community data exists but it is unstructured. Past events and their communities are documented by proceedings in digital libraries. Upcoming events are recognized by Call for Papers and detail information can be obtained from their web sites. There is no structured data for academic events. Moreover, with the recent advantages in technical communication as well as the increasing use of digital cooperation mechanism, there is also a requirement to integrate new digital media such as blogs, wikis, mailing-list, images, etc., into one model for events documentation. The model must reflect all aspects of events and their communities as well as be capable to connect and collect data from heterogeneous data sources such as digital libraries and the Web.

In this paper, we propose a model for events and scientific communities. Based on this model, we realize a SNA based approach to recommend the events to researchers. We study how the research communities support individual members in events finding by applying collaborative filtering technique for event

[5] `http://eventseer.net/`

recommendation. The paper is organized as follow. In the next section, we briefly survey the related work on Collaborative Filtering, Actor Network Theory and Social Network Analysis. In Section 3, we present a conceptual model for academic events and communities. In Section 4, the design of recommendation algorithm is discussed. In Section 5, we describe our experimental result with the real dataset from DBLP and EventSeer.net. In Section 6, we conclude our paper with a discussion and an outlook.

2 Related Work

Combining Social Network Analysis and recommender systems have been studied and applied in different application domains. In digital libraries, many approaches have been proposed to provide useful tools to researchers, e.g. citation recommendation [19], book recommendation [20], paper recommendation [21] etc. Generally, recommendation techniques can be categorized into three classes: Collaborative Filtering (CF), Content based and Hybrid approaches. CF is based on users community to generate recommendations, while Content based uses the features of items. Hybrid approaches combine CF and Content based with some other techniques such as demography, utility-based, knowledge-based recommendations to improve the quality of recommendation results. In this paper, we investigate how CF could be applied to event recommendation problem. We leave out hybrid approaches for the future work.

Collaborative Filtering (CF)
CF is widely used in commercial applications. CF provides the recommendations based on previous user's preferences and the opinions of other users who have similar preferences [4]. User's preferences can be expressed explicitly (e.g. rating for an item) or implicitly by interpreting user's behavior like purchase history, browsing data and other types of information access pattern. Collaborative filtering algorithms can be divided into two categories: memory-based collaborative filtering algorithms operate on the entire user-item database to generate the recommendations; model-based collaborative filtering algorithms use the user database to learn a model which is then used for recommending.

In general, a recommender system has three components: background data which is the information that the system has before the recommendation process begins, input data which is the information that user must communicate to the system in order to generate a recommendation, and an algorithm that combines background and input data to arrive at its suggestions [2]. In collaborative filtering, background data is the rating history of users on set of items, input data is rating history of target user. Collaborative filtering works by viewing the above dataset as a rating matrix. Ratings may be binary or real values indicate user's preference on the item. Columns in this matrix are items (called item vectors) and rows represent users (called user vectors). Each entry in the matrix is the user's rating for a particular item.

Actor Network Theory (ANT)
Actor-Network Theory (ANT) was developed by two French scholars, Michel Callon and Bruno Latour [7]. Digital networks are a meeting point for the social and technology. In ANT model, we have a network formulated by actors and relationships [8]. A actor may be a human or an object without any distinction. Any set of actors involved in a certain activity formulates a network. There are three special kinds of actors. The member stands for a person or a community. The medium enables members to do the activities, for example establishing communication links and exchanging the information. Artifacts are objects created by members using some media.

The conceptual model proposed in this paper is based on ANT. As mentioned earlier, digital media need to be integrated into the model for events and communities documentation. ANT tries to explain social order not through the notion of "the social" but through the networks of connections between human agents, technologies and objects [9]. Communities of academic events have been seen as communities of practice in which members exchange the information and communicate with each others using the combination of various communication methods such as face-to-face meeting and technology-enhanced methods, e.g. discussion forums, websites, mailing-list, blogs, wikis etc. Technology-enhanced communication techniques have became more and more important, especially when the international degree of recent conferences increases. Members of the community can live in different country and continents. Sometimes it is hard to organize face-to-face meeting and discussion. Therefore advance communication method is a important mechanism contributing to the successful of a scientific community. All these aspects need to be modeled as a cross-media base for scientific community.

Social Network Analysis
In digital library, it is possible to create the networks that reflects the collaboration between researchers using the references in research papers. In particular, there are many research work have studied the creation of these networks and applied Social Network Analysis for scientific community to understand the structure and pattern of research collaboration [10,11,12]. In the domain of publication and venue ranking, many approaches have been proposed to measure the impact of scientific collection (journals, proceedings) and scholar authors [15,16,17,18], which focuss on citation and co-authorship networks as the professional network between researchers. There are also researches which try to apply Social Network Analysis to evaluate the quality of academic events [6]. We are adding to these work by investigating the role of research community in helping researchers to find academic events and to identify research communities.

3 Model for Academic Events and Scientific Communities

Based on ANT, a model for academic events and communities is proposed as given in Figure 2. In this model, we consider the network of researchers in the relation with academic events. For each event (and event series), we have a network representing research collaboration between members. There are three kinds of network under consideration, including co-authorship network, citation network and co-participation network. *Scientist* entity describes the node of network, *Link* entity represents the connection between nodes and *Subnetwork* entity models the subnetwork extracted from *global network* which is composed of *Scientist* and *Link* entities. *Link* entity has a attribute *type* to differentiate three kinds of network: co-authorship, citation and co-participation networks.

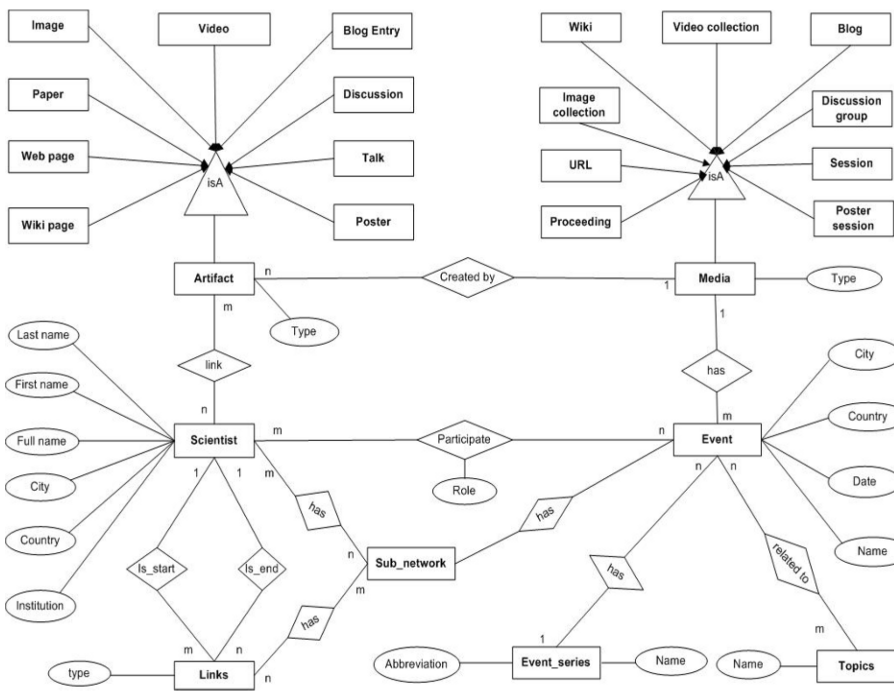

Fig. 2. Model for events and communities

Each *Event* belongs to a *Event series*, e.g. ACM SIGMOD, VLDB series etc. We consider all kinds of academic event, including conferences, workshops, international symposiums, doctoral consortiums as well as winter/summer schools. In general, workshops can be held as independent events (therefore they have their own series) or in combination with conferences, symposiums or consortiums. Each *Event* has a set of *Topics* which presents event's research domains and objectives. In fact, research topics tracking as well as topics classification are complicated problems. Research topics can be categorized in hierarchical structure in which a common topic (e.g. Database, Information Systems) can be divided into sub-classes. To keep it simple, in our model we use a "flat" list of topics used to specify research interests of a event. In mediabase, we integrate all types of digital media, e.g wikis, blogs, web sites, videos, images etc.

This model intends to be the basic on which a recommendation tool is based. It also serves as the foundation for event and community analysis. Mediabase could be extent so that different media management and monitoring tools like BlogWatchers, MailWatcher, WikiWatchers etc. can be applied.

4 Collaborative Filtering and Academic Event Recommendation

Standard Collaborative Filtering needs to be map to event recommending problem. In this section, we present a model and algorithm based on research communities of academic events. Formally, the problem can be stated as following:

Given a set of academic events E, set of researchers U and set of participation history vectors V in which $v_u = (e_1, e_2, ..., e_n)$ represents the participation history of researcher u. Recommend top K upcoming events for target researcher u_t.

General Algorithm

Standard collaborative filtering processes in three steps: building the model, computing similarity and generating recommendation. Our algorithm follows these steps and can be presented as following:

Input: set of events $E = (e_1, e_2, ..., e_N)$, set of researchers $U = (u_1, u_2, ..., u_M)$.

Output: top K most recommended events to target researcher u_t.

1. Building the model: construct the participating matrix $R(MxN)$.
2. Computing the similarity between target researcher u_t and others.
3. Generating recommendations: Select L most similar researchers and rank unknown events by aggregating the rating of L most similar researchers. Return most K ranked unknown events.

Building the Model

As presented in Section 2, collaborative filtering operates on a rating matrix in which each entry is user rating on an item. To map this model to our problem, we use the following approach: we consider academic events as "items" and researchers are users who will get the recommendations. The rating value of a researcher for an event is binary (i.e. 1 and 0), meaning that he participated or he will take part in this event, or not. We use event participation history of researchers as background data and the input data is the participation history of a particular researcher. The rating matrix then can be built using the background data. Formally, given a set of academic events E, set of researchers U then $R(M, N)$ is the rating matrix in which entry $R_{u,e} = 1$ if user u participated in event e and $R_{u,e} = 0$ if user u did not participated in event e. We use U_p to

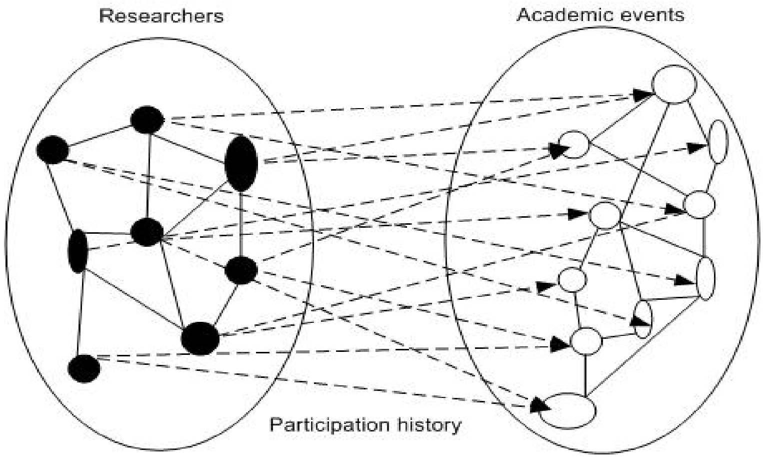

Fig. 3. Collaborative Filtering model mapping

denote the p^{th} row of R which is called the researcher vector of researcher u_p and E_q to denote q^{th} column of R which is called event vector of event e_q.

The above approach suffers from the *start-up* and *generality* problems. For newbie researchers who did not attend any events before nor publish any papers with other researchers, they will not get the recommendations since the system has no information about them. To overcome this problem, we use profile building mechanism as in other recommender systems: users have to rate a sufficient number of items before they can get the recommendations. Under the assumption that normally a newbie researcher starts his researches with the help of his professors or advisors as well as his colleagues. That means implicitly he has a research community. He could also join the communities of events in which he is interested, in order to keep track of what these communities is doing. Overall, by explicitly declaring his own "implicit" community, a newbie researcher can "embed" himself into a scientific community and let that community help him to find events.

Generality problem emerges from the fact that researchers may change their fields as well as work on different fields. For example, a researcher may work on database system and distributed system. Therefore, he attends conferences on database system and distributed system as well. Target researcher attended many conferences with him on database system and then he may be recommended conferences on distributed system. With many researchers like that, it is difficult to find a set of recommended events which satisfy target researcher's preferences. We solve this problem by a subjective classification via profile building mechanism. User's preference on topics is used to filter out events which are not relevant before performing recommendation process. This preprocessing procedure ensures that recommendation algorithm will work on a set of events which satisfies user's needs in general.

Computing Similarity

In this step we compute the similarity between researchers to find the set of most "closed" researchers to the target researcher. According to [5], various approaches can be applied to compute similarity. The two most popular approaches are *correlation* and *cosine-based*. In our work, we use *cosine-based* approach. To present them, let $E_{x,y}$ be the set of events which researcher x or researcher y, or both attended, i.e $E_{x,y} = \{e \in E \mid R_{x,e} = 1 \| R_{y,e} = 1\}$. $E_{x,y}$ is the union of events which researcher x and y attended (E_x and E_y relatively). In correlation approach, similarity function $sim(x, y)$ is computed by the Pearson correlation coefficient:

$$sim(x,y) = \frac{\sum_{e \in E_{x,y}} (R_{x,e} - \overline{R_x})(R_{y,e} - \overline{R_y})}{\sqrt{\sum_{e \in E_{x,y}} (R_{x,e} - \overline{R_x})^2 \sum_{e \in E_{x,y}} (R_{y,e} - \overline{R_y})^2}} \tag{1}$$

in which the average rating of researcher x, $\overline{R_x}$ is:

$$\overline{R_x} = \frac{1}{\mid E_x \mid} \sum_{e \in E_x} R_{x,e} \tag{2}$$

which is equals to 1 in our case.

In the cosine-based approach, the two researchers x and y are treated as two vectors $\overrightarrow{x}$ and $\overrightarrow{y}$ in m-dimensional space, where $m = \mid E_{x,y} \mid$. Similarity between two vectors can be measured by computing the cosine of the angle between them:

$$sim(x,y) = cos(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{\| \overrightarrow{x} \| \times \| \overrightarrow{y} \|} = \frac{\sum_{e \in E_{x,y}} R_{x,e} R_{y,e}}{\sqrt{\sum_{e \in E_{x,y}} R_{x,e}^2} \sqrt{\sum_{e \in E_{x,y}} R_{y,e}^2}} \tag{3}$$

where $\overrightarrow{x} \cdot \overrightarrow{y}$ denotes the dot-product between the vectors $\overrightarrow{x}$ and $\overrightarrow{y}$.

Generating Recommendations

Recommendation generating is a ranking process in which we compute a ranked values for unknown events. According to [5], ranked value is usually computed as an aggregate of the ratings of L most similar researchers for the same event:

$$R_{c,e} = aggr_{d \in C} R_{d,e} \tag{4}$$

where C denotes the set of L researchers who are most similar to researcher c and have participated in (or will attend) event e. Some of the aggregate functions are:

$$R_{c,e} = \frac{1}{L} \sum_{d \in C} R_{d,e} \tag{5}$$

$$R_{c,e} = k \sum_{d \in C} sim(c,d) \times R_{d,e} \tag{6}$$

$$R_{c,e} = \overline{R_c} + k \sum_{d \in C} sim(c,d) \times (R_{d,e} - \overline{R_d}) \tag{7}$$

where $\overline{R_c}$ is computed as in previous section and multiplier k serves as a normalizing factor and is usually selected as:

$$k = \frac{1}{\sum_{d \in C} sim(c, d)} \tag{8}$$

We use aggregate as an average (defined in the first case). However, in more complicated cases, the aggregate could be a weighted sum in which the similarity between c and d is used as a weight, i.e the more similar c and d are, the more weight $R_{d,e}$ will carry in the ranked value $R_{c,e}$.

5 Prototype Evaluation

Datasets

To evaluate the approach, a prototype is implemented based on the data from DBLP XML record and EventSeer.net[6]. First, DBLP XML record is parsed to get the list of past events and co-authorship network of each event. Event series are taken by parsing DBLP Website. Events then are bound into series by the unique URL prefixes of event and event series. Location information of events is also taken from DBLP Website. Upcoming events are extracted from EventSeer.net Web site. EventSeer.net contains most of Call for Papers for conferences in Computer Science. From EventSeer.net, we got a list of upcoming (and past) conferences with the information about time, locations, topics, persons and organizations. Overall, we have a dataset as summarized in Table 1.

Table 1. Dataset summary

Data	**Quantity**
Events	16821
Series	2099
Authors	522938
Topics	4910
Co-authorship of events	1282796 links

Data from DBLP and EventSeer.net is enough for the evaluation, although it is not complete. Ideally, we should have the list of all participants, authors and programm committee members (PC members) of each event. DBLP contains only the authors, while EventSeer.net indexes persons who are mentioned in Call for Papers, so they are mostly PC members. However, using authors and PC members as background data for recommendation algorithm is reasonable. Authors and PC members of each event have a closed relation via papers review process. Authors also have the knowledge about each others since they have worked on the same problems.

[6] `http://bosch.informatik.rwth-aachen.de:5080/AERCS/`

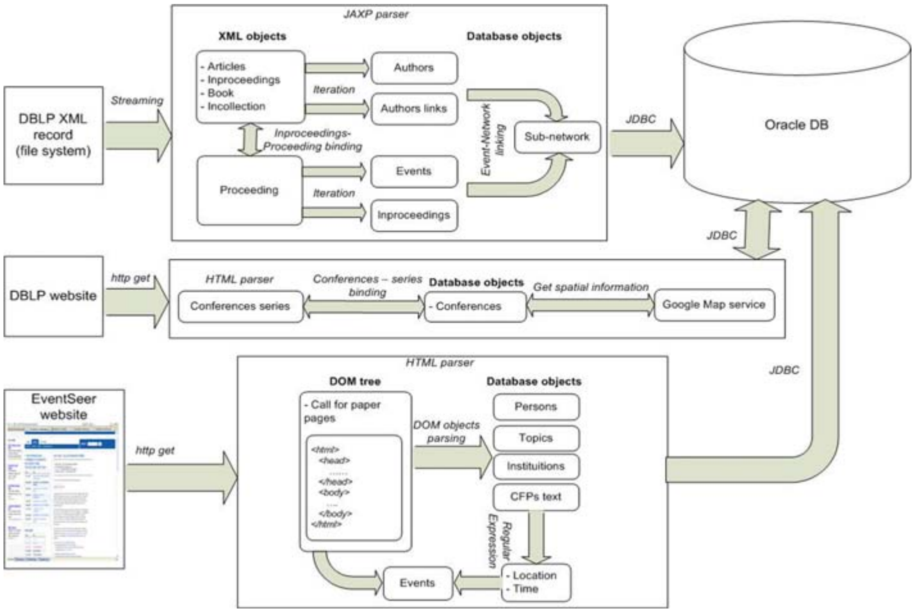

Fig. 4. Data preparation process

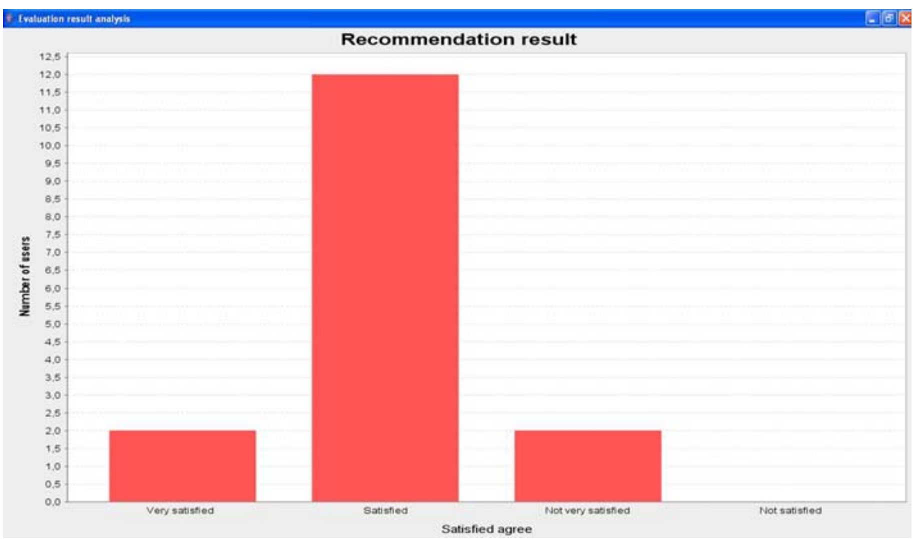

Fig. 5. Users satisfaction with recommendation result

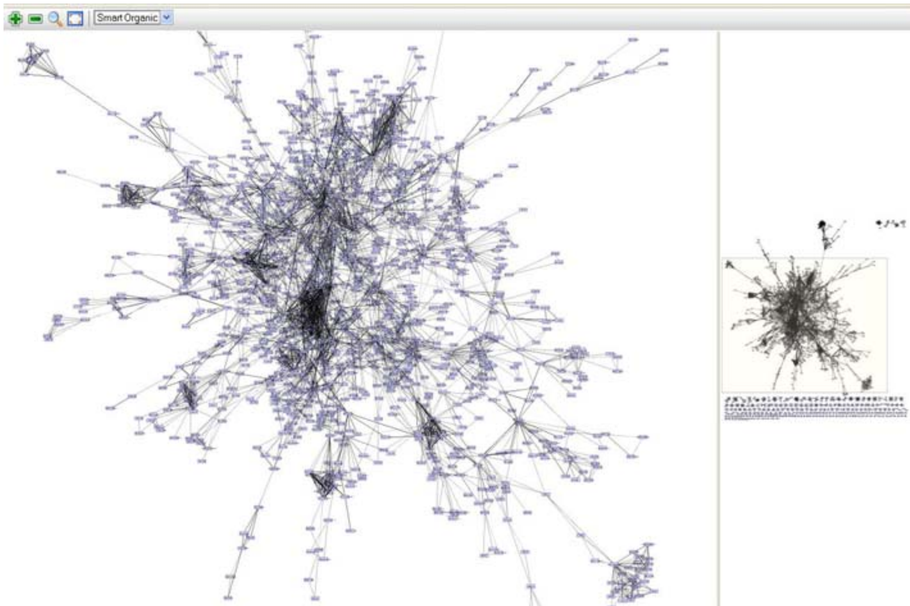

Fig. 6. ACM SIGMOD community visualization

Online Experiment

To evaluate the approach, we conducted a online survey on a set of users to get the opinion about recommendation result and community analysis provided by the system. Users are selected from colleagues and students working and studying at the Chair of Database and Information Systems, RWTH Aachen, Germany. A short tutorial is given to users and a questionaire is put online to let users answer a set of questions. The tutorial guides users through several tasks in order to get to know the concepts of the system, e.g. profile building, getting recommendation, finding events and event series as well as community analysis and visualization.

The system gains over 20 feedbacks in which most of the questions are filled in. First, we analyze the feedbacks to see users experiences in academic events as well as their roles in the events they attended. Most of users participated in 6 to 20 events, others attended 1 to 5 events. Among them, about 11 users took part in the events as participants, 6 users as presenters and a small number (about 4 users) as PC members. This result shows that our users community are young researchers.

In the second step, we assess the feedbacks to know users opinion about recommendation result. Users are asked to build their profiles in which they have to declare the preferences on topics, locations, persons and events. System then generates a list of upcoming events recommended to them. Users can compare the list with events they are interested in as well as discover new events which they do not know. Users express their opinion by answering a question about their satisfaction with recommendations. As shown in Figure 5, most of users satisfy with recommended events.

Besides recommending events to users, we perform the analysis on event series communities. This aims to provide to users a look inside the community of an event as well as event series. With our dataset, we are able to measure and present some parameters about the communities as proposed by Wenger et al. (2002) [14] and Kienle [13]. We analyze the development and continuity of communities by measuring the number of participants over years and number of participants according to the number of events they attended. Key members of the communities are also identified according to the number of events they attended in the series.

One of the most interesting features of the prototype is community visualization. We provide co-authorship network visualization of an event and event series as well as local network of a particular researcher. Community visualization is implemented based on yFile AJAX - a commercial network visualization tool. From the visualization, users can see the development of community of an event series over years as well as the community of event series as a whole.

6 Conclusions and Outlook

Recommender systems for digital libraries and scientific communities is an ongoing research domain. A recommender system could be a great tool for young researchers to find academic events to which they can submit papers. Our experiments show that applying a community based recommendation algorithm supports researchers in events finding. By using event participation history as background data for a Collaborative Filtering based algorithm, we are able to recommend the most relevant academic events to researchers. The algorithm works on the dataset which can be easily extracted from references in papers documented in digital libraries like DBLP, ACM or EventSeer.net.

The dataset of our system should be enhanced with some other data sources. Currently, data from DBLP and EventSeer.net is imported into our database. To have better recommendation results and analysis, we need also data from other digital libraries such as ACM, CiteSeer. The problem here is how to connect these data sources to provide a unique repository for academic events. We are working on this problem by investigating and applying different data and Web mining techniques in order to create a mesh data source network. Based on this, useful services could be further designed and implemented.

In the future, it would be interest to investigate other recommendation techniques as well as algorithms for event recommendation problem. Content-based recommendation and the combination of content-based with CF and other recommendation techniques is a promising direction. It would also be interesting to see these recommendation approaches in other domains. Currently, we are performing the social network analysis of 45.000 schools in Europe.

Another idea is to follow the dynamic behaviour of researchers. The movement of researchers between communities could be captured. The question is that what are important factors affecting this movement and the role of spanners in the communities. By tracking and analysis the dynamic movement of members, we could be able to recommend the future directions in research as well as carrier for researchers.

References

1. Schafer, J.B., Konstan, J.A., Riedl, J.: E-commerce recommendation applications. Data Min. Knowl. Discov. 5(1-2), 115–153 (2001)
2. Burke, R.: Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction 12(4), 331–370 (2002)
3. Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web, pp. 285–295. ACM Press, New York (2001)
4. Breese, J.S., Heckerman, D., Kadie, C.M.: Empirical analysis of predictive algorithms for collaborative filtering, pp. 43–52 (1998)
5. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the stateof-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (2005)
6. Zhuang, Z., Elmacioglu, E., Lee, D., Giles, C.L.: Measuring conference quality by mining program committee characteristics. In: Proceedings of the 2007 conference on Digital libraries, pp. 225–234. ACM Press, New York (2007)
7. Latour, B.: On recalling ant. In: Law, J., Hassard, J. (eds.) Actor-Network Theory and After, pp. 15–25 (1999)
8. Denev, D.: Multidimensional Patterns of Disturbance in Digital Social Networks. Master's thesis, RWTH Aachen University (2006)
9. Couldry, N.: Actor Network Theory and Media: Do They Connect and On What Terms? In: Hepp, A., et al. (eds.) Cultures of Connectivity. School of Economics and Political Science, London, pp. 1–14 (2004)
10. Newman, M.E.: Scientific collaboration networks. i. network construction and fundamental results. Phys. Rev. E. Stat. Nonlin. Soft. Matter. Phys. 64(1-2) (2001)
11. Newman, M.E.: Coauthorship networks and patterns of scientific collaboration. Proc. Natl. Acad. Sci. USA 101, 5200–5205 (2004)
12. Huang, T.H., Huang, M.L.: Analysis and visualization of co-authorship networks for understanding academic collaboration and knowledge domain of individual researchers. In: CGIV 2006: Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation, pp. 18–23. IEEE Computer Society, Washington (2006)
13. Kienle, A., Wesser, M.: Principles for cultivating scientific communities of practice. In: Proceedings of the 2nd International Conference on Communities and Technologies, pp. 283–299. Springer Netherlands (2005)
14. Wenger, E., McDermott, R., Snyder, W.M.: Cutivating Communities of Practice: A guid to Managing Knowledge. Havard Business School Press, Campridge (2002)
15. Yan, S., Lee, D.: Toward alternative measures for ranking venues: a case of database research community. In: Proceedings of the 2007 conference on Digital libraries, pp. 235–244. ACM Press, New York (2007)
16. Rahm, E., Thor, A.: Citation analysis of database publications. SIGMOD Record 34(4), 48–53 (2005)

17. Sidiropoulos, A., Manolopoulos, Y.: A citation-based system to assist prize awarding. SIGMOD Record 34(4), 54–60 (2005)
18. Sidiropoulos, A., Manolopoulos, Y.: A new perspective to automatically rank scientific conferences using digital libraries. Inf. Process. Manage. 41(2), 289–312 (2005)
19. McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., Riedl, J.: On the recommending of citations for research papers. In: Proceedings of the 2002 ACM conference on Computer supported cooperative work, pp. 116–125. ACM Press, New York (2002)
20. Mooney, R.J., Roy, L.: Content-based book recommending using learning for text categorization. In: Proceedings of the Fifth ACM Conference on Digital Libraries, pp. 195–204. ACM Press, New York (2000)
21. Torres, R., McNee, S.M., Abel, M., Konstan, J.A., Riedl, J.: Enhancing digital libraries with TechLens+. In: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 228–236. ACM Press, New York (2004)