Exploiting Social Networks for Publication Venue Recommendations

Hiep Luong[1], Tin Huynh[2], Susan Gauch[1] and Kiem Hoang[2]
[1]*University of Arkansas, Fayetteville, AR 72701, U.S.A.*
[2]*University of Information Technology, Ho Chi Minh City, Vietnam*

Keywords: Recommender Systems, Social Network Analysis, Publication History, kNN, Machine Learning.

Abstract: The impact of a publication venue is a major consideration for researchers and scholars when they are deciding where to publish their research results. By selecting the *right* conference or journal to which to submit a new paper minimizes the risk of wasting the long review time for a paper that is ultimately rejected. This task also helps to recommend appropriate conference venues of which authors may not be aware or to which colleagues often submit their papers. Traditional ways of scientific publication recommendation using content-based analysis have shown drawbacks due to mismatches caused by ambiguity in text comparisons and there is also much more to selecting an appropriate venue than just topical-matching. In our work, we are taking advantage of actual and interactive relationships within the academic community, as indicated by co-authorship, paper review or event co-organizing activities, to support the venue recommendation process. Specifically, we present a new social network-based approach that automatically finds appropriate publication venues for authors' research paper by exploring their network of related co-authors and other researchers in the same field. We also recommend appropriate publication venues to a specific user based on her relation with the program committee research activities and with others in her network who have similar paper submission preferences. This paper also presents more accurate and promising results of our social network-based in comparison with the baseline content-based approach. Our experiment, which was empirically tested over a large set of scientific papers published in 16 different ACM conferences, showed that analysing an academic social network would be useful for a variety of recommendation tasks including trend of publications, expert findings, and research collaborations, etc.

1 INTRODUCTION

The World Wide Web and its evolving infrastructure have played a significant role in the information explosion. This voluminous amount of unstructured and semi-structured data creates "big data," data sets that grow so large that they become awkward to work with or analyse using existing data management approaches. With the fast growth of digitalized textual data and documents, there is an urgent need for powerful text information management tools to help users find exactly what they are looking for and to help researchers keep abreast of information of whose existence they may be unaware. Recommender systems are one approach to helping users deal with the flood of information. They are tools that automatically filter a large set of items, e.g., movies, books, scientific papers, music, etc., in order to identify those that are most relevant to a user's interest.

There are a wide variety of dissemination outlets for research results, e.g., conferences, journals, seminars, scientific forums. When an author has a paper that they want to share, the review cycle can be time consuming and, if the paper is rejected because it is not a good fit, valuable time can be lost. In computer science, in particular, the pace of innovation is high. Selecting the right publication venue the first time is particularly important.

Traditional techniques usually use citation-based metrics with certain bibliometrics, such as the Impact Factor (Garfield, 1955), to measure the reputation and quality of publication. However, these techniques require frequent updates to the bibliometrics in order to maintain an accurate impact factor.

Recently there has been considerable interest in applying social network-based methods for ranking conference quality (Yan and Lee, 2007), seeking

Luong H., Huynh T., Gauch S. and Hoang K..
Exploiting Social Networks for Publication Venue Recommendations.
DOI: 10.5220/0004140102390245
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2012), pages 239-245
ISBN: 978-989-8565-29-7

research collaborators (Chen et al., 2011) or generating recommendations (Klamma et al., 2009); (Luong et al., 2012). Recommendation systems are particularly important for researchers and scholars in their professional research activities. For some experts in a research domain, or senior researchers who have strong publication records, selecting a conference might be a trivial task since they know well which conferences, journals, or scientific forums are the best places in which to publish their research papers. However, new researchers with less experience may not be able to easily assess conferences and the may not be current on relevant, new publication venues. The widespread use of the Internet allows researchers to create large, distributed academic social networks that can be analyzed to further enhance research productivity. Our interest is in how to use these academic social networks to recommend appropriate publication venues to authors for an unpublished paper.

In this paper, we present a survey of current research on content-based and collaborative filtering recommender systems, and recent trends applying social network analysis in recommender systems in section 2. We will focus mainly on recommendation research for academic activities and digital libraries. In section 3, we present our social network-based publication venues recommendation. Section 4 presents our content-based recommendation approach. Next, we present and discuss some experimental results for both appoaches in section 5. The final sections present our conclusions and discuss our future work in this area.

2 RELATED WORK

Traditional recommender systems are usually classified as content-based, collaborative, or hypbrid based on the type of information that they use and on how they use that information (Adomavicius and Tuzhilin, 2005). Content-based approaches compare the contents of the item to the contents of items in which the user has previously shown interest. Automated text categorization is considered the core of content-based recommendation systems. (Yang et al., 1999) reported a controlled study with statistical significance tests on five text categorization methods: Support Vector Machines (SVM), k-Nearest Neighbors (kNN) classifier, neural network approach, Linear Least-squares Fit mapping and a Naïve Bayes classifier. Their experiments with the Reuters data set showed that SVM and kNN significantly outperform the other classifiers, while Naïve Bayes underperforms all the other classifiers. In other work, kNN was found to be an effective and easy to implement that could, with appropriate feature selection and weighting, outperform SVM (Cunningham and Delany, 2007).

Collaborative Filtering (CF) determines similarity based on collective user-item interactions, rather than on any explicit content of the items. (Su and Khoshgoftaar, 2009) has summarized a detail review of some main CF recommendation techniques. In another recommendation research using CF.

The online world has supported the creation of many research-focused digital libraries such as the Web of Science, ACM Portal, Springer Link, IEEE Xplore, Google Scholar, and CiteSeerX. Recently, new research is focusing on these as enablers of a community of scholars, building and analyzing social networks of researchers to extract useful information about research domains, user behaviours, and the relationships between individual researchers and the community as a whole. Microsoft Academic Search (MAS), ArNetMiner (Tang et al., 2008), and AcaSoNet (Abbasi and Altmann, 2011) are online, web-based systems whose goal is to identify and support communities of scholars via their publications. The entire field of social network systems for the academic community is growing quickly, as evidenced by the number of other approaches being investigated (Abbasi and Altmann, 2011); (Miki et al., 2005) and (Mika, 2005).

In order to extracting useful information from an academic social network, (Zhuang et al., 2007) proposed a set of novel heuristics to automatically discover prestigious (and low quality) conferences by mining the characteristics of Program Committee members. (Chen et al., 2011) introduces CollabSeer, a system that considers both the structure of a co-author network and an author's research interests for collaborator recommendation. CollabSeer suggests a different list of collaborators to different users by considering their position in the co-authoring network structure. In work related to publication venues recommendation, (Pham et al., 2011) proposed a clustering approach based on the social information of users to derive the recommendations.

To our best knowledge, we have not seen existing research that exploits the relationships between the authors of an unpublished paper with conference PC members or people in the social network who have previous publications to recommend appropriate publication venues.

3 SOCIAL NETWORK-BASED RECOMMENDATION APPROACH

3.1 Overview

We introduce a new approach to recommending a list of appropriate conference venues to an author for their unpublished paper by using a large-scale network of researchers. By analysing this large-scale social network, we recommend publication venues to the unpublished paper's authors based on the 'similarity' they have with conference PC members or with other authors with papers published in the conferences.

To build our dataset, we selected four subdomains of research in Computer Science corresponding to four SIGs (Special Interest Groups). We have chosen four different fields to challenge our recommendation task. Then, we manually picked four different conferences for each SIG. The total list of 16 selected conferences was presented in our previous work (Luong et al., 2012). For each of these 16 conferences, we downloaded all published papers from 2008-2010 from the ACM digital library as well as the list of Program Committee (PC) members for each conference. We have built a truth-list of correct paper-conference relation to evaluate our conference recommendation approaches.

3.2 Academic Social Network Analysis

We measure the closeness between authors and conferences by two new methods, the first based on the relationships between the candidate paper authors and the PC members, the second between the candidate paper authors and previously authors previously published in each conference.

This problem can be formalised as a relatedness calculation of vertices in a graph. Each researcher can be viewed as a vertex of the graph, and the graph edges represent co-author relationships with other researchers in the network. Note that a member (a vertex) in the network can be either a paper author, a PC member, or both. Figure 1 represents part of an academic social network in which A_1 and A_2 are seeking an appropriate publication venue for their jointly authored, unpublished paper. In this figure, the blue nodes represent the candidate paper authors, the white nodes other researchers, and the orange nodes represent a particular conference's PC member. In our initial work, we weight all edges with 1 for simplicity.

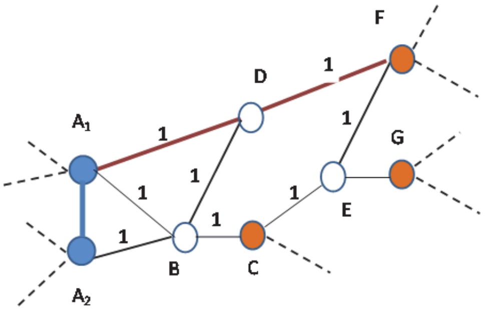

Figure 1: Example of a part of the social network.

The closeness between authors is calculated based on the shortest path connecting them. For example, to determine the relation similarity between author A_1 and a PC member F, there are several paths between these two nodes such as A_1-D-F, A_1-B-D-F, A_1-B-C-E-F, etc. However, their shortest path, A_1-D-F, contains 2 edges. The shorter the path, the closer the nodes are. Thus, closeness between two nodes in a graph, A and B, is the inverse of path length calculated as:

$$Closeness(A, B) = \frac{1}{ShortestPath(A, B)}$$

3.2.1 Closeness between Paper Authors and PC Members (*Author_PC* Method)

In this method, we recommend that a paper be submitted to the conference with the strongest relationship between the paper's authors and the conference's program committee. Essentially, this measures the impact that being an insider has to the likelihood of a paper being accepted. The more close relationships that a paper's authors have with PC members, the more likely that conference is to be recommended for the paper. Note that this favors conferences with large program committees and, in future work, we will investigate the effects of normalizing these weights by the number of PC members.

In particular, for conference *i*, we add together the strength of the relationship between each of the paper's co-authors with the PC as follows:

$$Author\_PC_i = \sum_{m=1}^{A} Closeness(m, i)$$

where *A* is the number of authors of the unpublished paper and *Closeness(m,i)* is the closeness between an author *m* and the program committee for conference *i*. Specifically, *Closeness(m, i)* is the sum of the closeness between that author and any member of the PC.

$$Closeness(m,i) = \sum_{j=1}^{N} Closeness(m, PC_{i,j})$$

with *m* is an author of the paper, $PC_{i,j}$ is the j^{th} program committee member for conference *i*, and *N* is the total number of PC members for that conference. The conference with the highest value is recommended as the publication venue.

3.2.2 Closeness between Paper Authors and Previous Conference Authors (*Author_NetAuthors* Method)

In this method, we consider how close a paper's authors are to those in the network who have published their paper(s) in a specific conference. Essentially, this is based on the belief that, if papers authors academic colleagues have had their work published by a particular conference in the past, this is a good indication that this paper is also likely to be acceptable. This relation similarity can be defined as following:

$$Author\_NetAuthors_i = \sum_{m=1}^{A} Closeness(m, NetAuthors_i)$$

where *A* is a number of author(s) of the unpublished paper, and *Closeness(m, NetAuthors$_i$)* is the closeness between an author *m* and all other authors in the network who have published their papers to the conference *i*. This value *Closeness(m, NetAuthors$_i$)* is also calculated as the sum of the closeness between that author with any other author relevant to the conference *i*.

$$Closeness(m, NetAuthors_i) = \sum_{j=1}^{M} Closeness(m, Author_{i,j})$$

with *m* is an author of the paper, *Author$_{i,j}$* is the j^{th} author in the network who has paper published in the conference *i*. *M* is total number of all members that have papers published in the conference *i*. We recommend the conference with the highest value as a publication venue for the unpublished paper.

4 EXPERIMENTS

4.1 Dataset

In order to get the publication history of authors, we developed a focused crawler in Java that extracts all co-authors and relevant publications for a given author from the Microsoft Academic Search (MAS) website. As presented in the section 3.1, our input contains 16 ACM conferences of 4 SIGs. With all papers collected from 2008-2010, we used the papers published in two years 2008 and 2009 as training documents and the ones published in 2010 as test documents for the classification task. Since the number of papers published for each conference varies, we randomly selected 20 documents per year per conference (60 totals). Thus, each conference had 40 training and 20 test documents. With 16 conferences, the total test collection contained 640 training and 320 test documents. We split the 320 test documents into two sets: 160 for tuning and 160 for validation. For each of the 16 conference instances, we also downloaded the names of the program committee members from the conference website for the year 2010.

For each paper in the test collection, we extracted the author names and used a crawler to gather information about each author's publications and co-authors. The co-authors of the co-authors were then recursively collected, until a network 3 levels deep was created. As a result, we collected information about 306,227 authors and 392,878 papers. We also submitted the names of the PC members to MAS to collect their authorship relationship. This information was downloaded and stored in a database. We manually reviewed authors with large numbers of publications to remove publications that were incorrectly attributed to an author due to the ambiguity of the author's name.

Finally, we built a large-scale graph, representing the network of researchers and experts, containing 303,843 vertices (authors) and 1,220,472 edges (co-authorship relationships with an average node degree of 8.03.

4.2 Baseline

In order to evaluate our new recommendation approach using academic social networks, we compared the methods described in section 3 with a baseline of content-based recommendations. In the baseline, conferences in the dataset are treated as different categories into which we classify a candidate paper. Each paper is represented by a vector of terms weighted by TF-IDF. The similarity between two papers is calculated using the cosine similarity measure. Papers are classified into conferences using a k-Nearest Neighborhood (kNN) algorithm (Gauch et al., 2004) trained using the 640 training documents. In order to identify the best k value for our experiment, we varied k, the number of classifier results used for to select the conference, from 3-40. Based in previous testing (Luong et al., 2012), we report our best performing k (i.e., k = 25) at which the content-based approach reached the highest classification precision. We also compare the performance of our new research to our previous work, i.e., PubHistory that recommends conferences to authors based on their own publication histories, rather than the publication histories of researchers in their academic social network (Luong et al., 2012).

4.3 Evaluation Metrics

We evaluated each method's performance using precision, i.e., the percentage of the time that the recommender system recommends the conference and/or SIG in which the test paper actually appeared:

- *Conference Precision*: measures the percentage of time that the recommender recommends the true conference. This is reported at various cut-offs, i.e., Top1 means that the correct conference was the top-ranked recommendation etc.
- *SIG Precision*: measures the percentage of time that the recommender recommends a conference from the correct SIG. Since there are fewer SIGS, and they differ more than the conferences do, this is an easier task.

4.4 Results and Discussion

Table 1 summarizes the conference precision results for each of the four methods using the 160 tuning documents. The conference precision results are also shown graphically in Figure 2.

Table 1: Conference precision results using four different methods.

	Conference Precision			
	Top1	**Top2**	**Top3**	**Top4**
Content-Based	*48.8%*	*68.8%*	*80.6%*	*90.6%*
PubHistory	66.2%	83.8%	95.5%	96.8%
Author_PC	43.6%	59.6%	69.9%	77.6%
Author_Net Authors	**74.5%**	**91.0%**	**94.3%**	**96.8%**

These results show that the Author_NetAuthors method is by far the most accurate method for recommending a conference. It recommends the correct conference as the top choice 74.5% of the time and the correct conference is within the top 4 choices almost 97% of the time. Authors tend to submit to, and be accepted by, conferences in which their co-authors (direct or indirect) have previously been published. Interestingly, this method outperforms the PubHistory method (66.2% Top1 conference precision) that recommends conferences based only on the author's own publication history. The difference is most evident in the Top1 and Top2 conference precision. Clearly, information from the author's academic social network helps identify good publication venues.

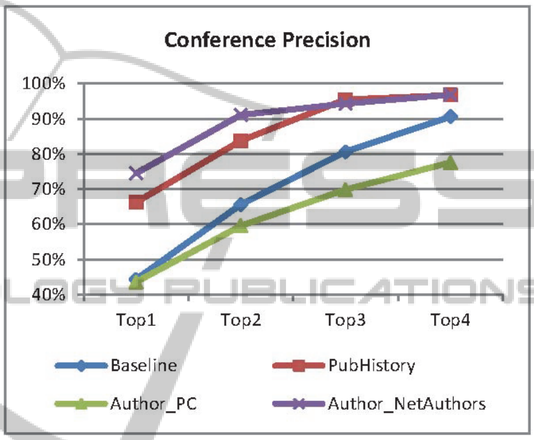

Figure 2: Comparison of classification precisions.

The two methods based on publication history, Author_NetAuthors and PubHistory, both outperform the _content-based baseline method by a wide margin. The content-based baseline achieved a Top1 conference precision of only 49%. The Author_PC method (43.6% Top1 conference precision) is the only method that underperforms the Content-Based method. This is actually a very positive result for the integrity of the research community. It demonstrates that relationships between authors and PC members does not predict acceptance of a paper at a conference. It is more important that the paper be of interest to the community, and on topic, than that the author has worked with someone on the PC in the past.

Table 2: SIG precision using four different methods.

Methods	**SIG Precision for the top-ranked**
Content-Based	80.63%
PubHistory	87.66%
Author_PC	74.36%
Author_NetAuthors	93.00%

Since there are 16 conferences that overlap in topics, but only 4 SIGs that cover different research areas, predicting the correct SIG should be an easier task than predicting the correct conference. Table 2, which summarizes the SIG precision results for the top-ranked result of four methods, confirms this hypothesis. These results confirm our hypothesis that a publication venue recommendation system can benefit from social network analysis instead of, or in addition to, traditional content-based approaches.

6 CONCLUSIONS

The goal of this research is to implement and evaluate a new approach to recommend publication venues for an unpublished article. Our approach takes advantage of information analysed from an academic social network of researchers linked by their co-authorship relationships. The results show that the Author_NetAuthors approach that incorporates relationships between a paper's authors' academic social network and each conference's network of previously published authors is the best performing result. Overall, we conclude that social network-based approaches can outperform content-based approaches when recommending publication venues. They work well even when deciding between conferences that overlap in topics, a task that is very difficult for content-based recommender systems. We also showed that relationships with the community of authors who publish in specific conferences is more important than relationships with members of the conference's program committee members.

Our main tasks in the future are to enhance the publication venue recommendation system by developing algorithms that take into account more sophisticated graph relationships and different kinds of links in the network such as citation and other indications of research collaboration (e.g., researchers from the same institution).

ACKNOWLEDGEMENTS

This research is partially supported by the NSF grant number 0958123 - Collaborative Research: CI-ADDO-EN: Semantic CiteSeer^X.

REFERENCES

Abbasi, A. and Altmann, J. 2011. "On the Correlation between Research Performance and Social Network Analysis Measures Applied to Research Collaboration Networks," *44th Hawaii International Conference on System Sciences*, 2011, pp.1-10.

Adomavicius, G., and Tuzhilin, A. 2005. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." *IEEE Trans. On Knowledge and Data Engineering*, pp. 734-749.

Cunningham, P., Delany, S.J. 2007. "k-Nearest Neighbour Classifiers" In: *Multiple Classifier Systems*, pp. 1--17. University College Dublin (2007)

Garfield, E. 2006. "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas". *Int. J. Epidemiol.* 35 (5): pp. 1123-1127.

Gauch, S., Madrid, J. M., Induri, S., Ravindran, D. and Chadlavada, S. 2004. "*KeyConcept: A Conceptual Search Engine*", Center, Technical Report: ITTC-FY2004-TR-8646-37, University of Kansas.

Hung-Hsuan Chen, Liang Gou, Xiaolong Zhang, and Clyde Lee Giles. 2011. "CollabSeer: a search engine for collaboration discovery." In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries (JCDL '11)*. NY, USA, 231-240.

Klamma, R. & Pham, M. C., & Cao, Y. (2009). "You Never Walk Alone: Recommending Academic Events Based on Social Network Analysis." *Proceedings of the First International Conference on Complex Science (Complex'09)*, Feb. 23-25, 2009, China.

Luong, H., Huynh, T., Gauch, S., Do, L., and Hoang, K. 2012. "Publication Venues Recommendation using Author Networks Publication History", *The 4th Asian Conference on Intelligent Information and Database Systems*, March 19-21, 2012, Kaohsiung, Taiwan. ACIIDS (3) 2012: pp.426-435.

Microsoft Academic Search, http://academic.research.microsoft.com

Mika, P., 2005. "Flink: Semantic web technology for the extraction and analysis of social networks." *Journal of Web Semantics*, vol. 3(2).

Miki, T., Nomura, S., Ishida, T., 2005. "Semantic Web Link Analysis to Discover Social Relationships in Academic Communities." *IEEE Computer Society*, pp. 38-45.

Pham, M. C., Cao, Y., Klamma, R. and Jarke, M. 2011. "A clustering approach for collaborative filtering recommendation using social network analysis*." J. UCS*, 17(4):583-604, 2011.

Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Adv. in Artificial Intelligence*, 2009:1–20.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z. 2008. "Arnetminer: extraction and mining of academic social networks." *Proceeding of the 14th ACM SIGKDD*. NY, USA (2008), pp. 990-998.

Yan, S., & Lee D. (2007). "Toward alternative measures for ranking venues: a case of database research community." *Proc. of the 7th ACM/IEEE joint conference on digital libraries*, pp. 235–244.

Yang, Y., Liu, X. 1999. "A re-examination of text categorization methods." *Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 42--49. Berkeley, California, United States (1999)

Zhuang, Z., Elmacioglu, E., Lee, D., & Giles, C. L. (2007). "Measuring conference quality by mining program committee characteristics." *Proceedings of the 7th ACM/IEEE joint conference on digital libraries, ACM*, New York, USA, pp. 225–234.