A Topic Modeling Approach for Research Community Mining

Ali Daud
Department of Computer Science
Sector H-10, International Islamic University, Islamabad, 44000, Pakistan
ali.daud@iiu.edu.pk

Abstract

Mining community on the basis of hidden relationships present between the entities is important from academic recommendation point of view. Previous approaches mined research community by using network connectivity or by ignoring semantics-based intrinsic structure of the words and author's relationships present between the conferences. In this paper, we propose a novel Venue-Author-Topic (VAT) approach which can consider semantics-based intrinsic structure of words and authors correlations, simultaneously. We also show how topics and authors can be inferred for new conferences and authors correlations can be discovered by using proposed approach. Experimental results on the corpus downloaded from DBLP shows the effectiveness of proposed approach and the detailed interpretation of results reveals interesting information about the research community.

Keywords
Semantic Analysis, Digital Libraries, Community Mining, Unsupervised Learning

1. Introduction

Community mining by exploiting the relationship between entities is an active area of research. For example, various conferences are held every year about different topics and huge volume of scientific literature is collected about conferences in the digital libraries. It provides us many challenging discovery tasks which are very useful from academic recommendation perspective. For example, to find reviewers for a specific area of conference, suggesting conferences to the researchers for submitting papers, inviting program committee members for a conference etc.

Previously, research community mining (by which we mean related authors and conferences) problem is investigated by considering network connectivity [20,21] or by semantics-based intrinsic structure of the words (topic-based retrieval) [13] without considering conferences information. Consequently, [17] proposed a topic-based approach which can discover research community by considering conferences information. They viewed conference information just as a stamp (token), which became the reason of ignoring implicit semantics-based words and authors correlations present between the conferences. Previous approaches do not investigate the latent topics of the conferences explicitly, by ignoring conferences internal semantic text dependencies and authors correlations. While, in real world conferences internal topics and author correlations are very important for finding specific research area conferences and matching reviewers with papers.

In this paper, we investigate the problem of research community mining by modeling conferences latent topics and authors together. We generalized previous topic modeling approach [17] form a single document to all publications of a conference, which can provide grouping of conferences and authors in different groups on the basis of latent topics present in conferences. We propose a Venue-Author-Topic (VAT) approach a variation of Author-Topic model [16] which can discover research community and can be used to find correlations between authors. We can say that solution provided by us for research community mining problem produced quite intuitive, realistic and functional results.

The novelty of work described in this paper lies in the formalization of the research community mining problem from conference level (CL), generalization of previous topic modeling approach from document level to CL (VAT), and experimental verification of the effectiveness of our proposed approach on the real-world corpus. To the best of our knowledge, we are the first to deal with the aforementioned research community mining problem by proposing a generalized topic modeling approach from document level (DL) to conference level (CL).

The rest of the paper is organized as follows. In Section 2, we formalize the problem. Section 3 illustrates our proposed approach for modeling research community with its parameter estimation details. In Section 4, corpus, experimental setup with empirical studies and discussions about the results are given. Section 5 provides related work and section 6 brings this paper to conclusions.

2. Problem Setting

Automatic discovery of hidden associations present between entities from digital libraries is very useful. Our work is focused on mining research

community by modeling the relationships between its entities on the basis of semantics-based text information. Each conference accepts many papers every year written by different authors. To our interest, each publication contains some title words and names, which usually covers most of the highly related sub research areas of conferences and authors, respectively. Conferences with their accepted papers on the basis of latent topics can help us to discover research community.

We denote a conference c as a vector of N_c words based on the paper titles accepted by the conference, an author r as an attendee of the conference on the basis of his accepted paper (s), and formulize research community mining problem as: Given a conference c with N_c words, and $\mathbf{a}_c$ authors of a conference c, discover semantically related research community entities (e.g. topically related authors and conferences). To perform semantic analysis of research community our proposed approach can smooth data from semantic level by considering intrinsic text dependencies.

3. Research Community Mining

In this section, before describing our VAT approach, we will first describe how documents and authors are modeled with topics.

3.1. Modeling Documents with Topics

Topic modeling which can capture the semantics-based structure of words, assumes that there is a hidden topic layer $T = \{z_1, z_2, z_3, \ldots, z_t\}$ between the word tokens and documents, where z_i denotes a latent topic and each document d is a vector of N_d words $\mathbf{w}_d$. A collection of D documents is defined by $D = \{\mathbf{w_1}, \mathbf{w_2}, \mathbf{w_3}, \ldots, \mathbf{w_d}\}$ and each word w_{id} is chosen from a vocabulary of size V. For each document, a topic mixture distribution is sampled and a latent topic T is chosen with the probability of topic given document for each word with word having generated probability of word given topic [5,12].

3.2. Modeling Authors with Topics

Following basic idea of topic modeling words and authors are modeled by considering latent topics to discover the research interests of authors using Author-Topic model (AT) [16]. In AT, each topic is associated with a multinomial distribution Φ_z over words of a document. Each author (from a set of K authors) is associated with a multinomial distribution θ_a over topics. Both θ_a and Φ_z have symmetric Dirichlet prior with hyper parameters α and β. For each word in the document, an author r is uniformly sampled from set of coauthors $\mathbf{a}_d$, then topic z is sampled from the multinomial distribution θ_a associated with author r and word w is sampled from multinomial topic distribution Φ_z associated with topic z.

3.3. Venue-Author-Topic Approach (VAT)

The basic idea presented in AT [16], that words and authors of documents can be modeled by considering latent topics became the intuition of modeling words, authors and conferences, simultaneously to exploit hidden research community. In VAT, we viewed a conference as a composition of documents words and the authors of its accepted publications. Symbolically, for a conference c we can write it as: $C = \{(\mathbf{d}_1,\mathbf{a}_{d1})+(\mathbf{d}_2,\mathbf{a}_{d2})+(\mathbf{d}_3,\mathbf{a}_{d3})+ \ldots + (\mathbf{d}_i,\mathbf{a}_{di})\}$, where d_i is a document of a conference and a_{di} are authors of document d_i.

DL approaches consider that an author is responsible for generating some latent topics of the documents. While, our CL approach considers that an author is responsible for generating some latent topics of the conferences. In VAT, each topic is associated with a multinomial distribution Φ_z over words of a conference. Each author (from set of K authors) of a conference c is associated with a multinomial distribution θ_a over topics. Both θ_a and Φ_z have symmetric Dirichlet prior with hyper parameters α and β. For each word in the conference, an author r is uniformly sampled from set of author's $\mathbf{a}_c$, then topic z is sampled from the multinomial distribution θ_a associated with author r and word w is sampled from multinomial topic distribution Φ_z associated with topic z.

The generative process is as follows:

1. For each topic $z = 1,\ldots, T$
   Choose Φ_z from Dirichlet (β)
2. For each author $r = 1,\ldots, K$ of conference c
   Choose θ_a from Dirichlet (α)
3. For each word $w = 1,\ldots, N_c$ of conference c
   Choose an author r uniformly from all authors $\mathbf{a}_c$
   Choose a topic z from multinomial (θ_a) conditioned on r
   Choose a word w from multinomial (Φ_z) conditioned on z

We utilize Gibbs sampling [1] for parameter estimation in our approach which has two latent variables z and r; the conditional posterior distribution for z and r is given by:

$$P(z_i = j, r_i = k | w_i = m, \mathbf{z}_{-i}, \mathbf{r}_{-i}, \mathbf{a}_c) \propto \frac{n_{-i,j}^{(wi)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(ri)} + \alpha}{n_{-i,\cdot}^{(ri)} + T\alpha} \quad (1)$$

where $z_i = j$ and $r_i = k$ represent the assignments of the i^{th} word in a conference to a topic j and author k respectively, $w_i = m$ represents the observation that i^{th} word is the m^{th} word in the lexicon, and z_{-i} and r_{-i} represents all topic and author assignments not

including the i^{th} word. Furthermore, $n_{-i,j}^{(wi)}$ is the total number of words associated with topic *j*, excluding the current instance, and $n_{-i,j}^{(ri)}$ is the number of times author *k* is assigned to topic *j*, excluding the current instance, and *W* is the size of the lexicon. "." indicates summing over the column where it occurs and $n_{-i,j}^{(\cdot)}$ stands for number of all words that are assigned to topic *z* excluding the current instance.

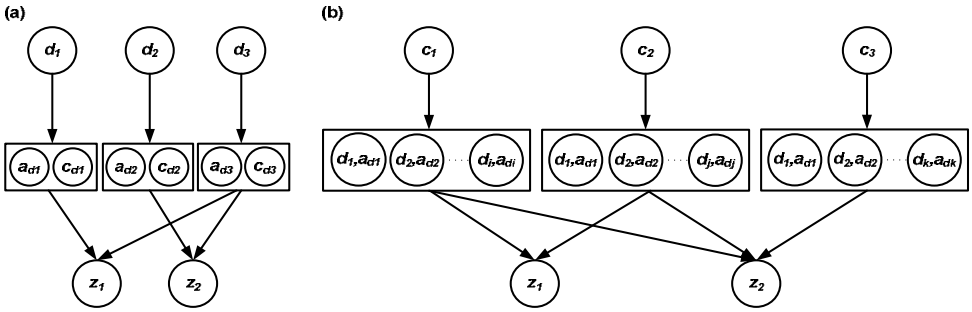

Figure 1: Research community topic modeling (a) DL and (b) CL approaches.

During parameter estimation, the algorithm only needs to keep track of *W* x *T* (word by topic) and *T* x *R* (topic by author) count matrices. From these count matrices, topic-word distribution Φ and conference-topic distribution θ can be calculated as:

$$\emptyset_{zw} = \frac{n_{-i,j}^{(wi)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \quad (2)$$

$$\theta_{rz} = \frac{n_{-i,j}^{(ri)} + \alpha}{n_{-i,\cdot}^{(ri)} + T\alpha} \quad (3)$$

where, $\emptyset_{zw}$ is the probability of word *w* in topic *z* and θ_{rz} is the probability of topic *z* for author *k*. These values correspond to the predictive distributions over new words *w* and new topics *z* conditioned on *w* and *z*. To find *T* x *C* (topic by conference) count matrix we calculated the probability distribution of topic given conference as given below, where, $\mathbf{a}_c$ is the number of authors belongs to a conference *c*.

$$p(z|c) = \sum_{a \epsilon A_c} p(z|a)p(a|c) = \frac{1}{|A_c|} \sum_{a \epsilon A_c} p(z|a) \quad (4)$$

4. Experiments

4.1. Corpus

We downloaded five years publication corpus of conferences from DBLP [6] for years 2003-2007. In total, we extracted 112,317 authors, 62,563 publications, and combined them into a single document separately for 261 conferences. We then processed corpus by (a) removing stop-words, punctuations and numbers (b) down-casing the obtained words of publications, and (c) removing words and authors that appear less than three times in the corpus for usual preprocessing done for text mining. This led to a vocabulary size of V=10,872, a total of 572,592 words and 26,078 authors in the corpus.

4.2. Experimental Settings

We use Gibbs sampling algorithm [11] for finding optimal values of hyper-parameters α *and* β. In our experiments, for 150 topics *T* the hyper-parameters α and β were set at 50/T and 0.01 respectively, by following the values used in [16]. The number of topics *T* was fixed at 150 on the basis of human judgment of meaningful topics and measured perplexity [2] on 20% held out test corpus for different number of topics *T* from 2 to 300.

4.3. Results and Discussions

4.3.1. Conferences, Authors and Topics. Authors and conferences related to specific area of research on the basis of latent topics are extracted. Table1 illustrates 4 different topics out of 150, discovered from the 120th iteration of the particular Gibbs sampler run. Each topic shows 8 words that are most likely to be produced, and the 8 authors and conferences that are most likely to be related to words in a specific topic.

The words, authors and conferences associated with each topic are also quite representative as they show semantic based community of a specific area of research. Here it is obligatory to mention that top 8 authors and conferences associated with a topic are not necessarily the most well-known authors and

conferences in that area, but rather are the authors and conferences that tend to produce most words for that topic in the corpus and are responsible for creating hidden research community. But one can see that, top ranked discovered authors and conferences for different topics are typically experts and top class conferences of that area of research, respectively. For example, in case of topic 11 "Information Retrieval" and topic 54 "XML Databases" top ranked authors and conferences are well known in their respective fields.

The topic # 54 "XML Databases" and topic # 100 "Software Engineering" shows quite specific and precise topics with move from simple databases to XML database and high importance of agile processing for software maintenance in software engineering, respectively. Topics # 99, 54, 11 are topics with direct relevance to data mining and Web namely, data mining itself, XML databases and information retrieval.

Proposed approach discovers several other topics related to data mining such as neural networks, multi-agent systems and pattern matching, also other topics that span the full range of areas encompassed in the corpus.

Topic 54 "XML Databases"		**Topic 100 "Software Engineering"**		**Topic 11 "Information Retrieval"**		**Topic 99 "Data Mining"**	
Word	***Prob.***	***Word***	***Prob.***	***Word***	***Prob.***	***Word***	***Prob.***
Xml	0.093811	Software	0.197374	Retrieval	0.127699	Mining	0.15788
Data	0.079758	Development	0.059311	Text	0.092182	Clustering	0.10328
Query	0.072938	Engineering	0.053288	Query	0.05448	Data	0.070152
Queries	0.053479	Component	0.035498	Relevance	0.037277	Classification	0.049702
Databases	0.051068	Testing	0.032255	Information	0.029392	Patterns	0.037126
Database	0.042586	Agile	0.027159	Search	0.023583	Frequent	0.028128
Processing	0.026536	Test	0.026695	User	0.020074	Discovery	0.027515
Relational	0.025205	Requirements	0.026047	Language	0.018924	Association	0.021993
Author	***Prob.***	***Author***	***Prob.***	***Author***	***Prob.***	***Author***	***Prob.***
Wei Wang	0.011326	Frank Maurer	0.013063	Alexander F. Gelbukh	0.011602	Philip S. Yu	0.021991
Divesh Srivastava	0.010205	Mario Piattini	0.009204	W. Bruce Croft	0.010554	Reda Alhajj	0.01785
Elke A. Rundensteiner	0.009747	Baowen Xu	0.007954	Wei-Ying Ma	0.010379	Jiawei Han	0.017792
Kian-Lee Tan	0.008881	Steacute Ducasse	0.006976	Barry Smyth	0.010205	Hans-Peter Kriegel	0.014686
Sourav S. Bhowmick	0.008779	John C. Grundy	0.006812	Zheng Chen	0.009855	Wei Wang	0.01066
Divyakant Agrawal	0.008269	Grigori Melnik	0.006323	ChengXiang Zhai	0.009739	Eamonn J. Keogh	0.010373
Nick Koudas	0.008218	Gerardo Canfora	0.006269	Chew Lim Tan	0.009623	Christos Faloutso	0.009567
Gerhard Weikum	0.007913	Arie van Deursen	0.005399	Maarten de Rijke	0.007702	Ming-Syan Chen	0.008992
Conference	***Prob.***	***Conference***	***Prob.***	***Conference***	***Prob.***	***Conference***	***Prob.***
Xsym	0.094329	Agile Deve.	0.104955	TSD	0.050347	DAWAK	0.043478
SSDBM	0.06592	XP	0.076159	ECIR	0.048561	SDM	0.042625
ADBIS	0.050305	WCRE	0.058275	ACL	0.045584	PKDD	0.03843
ADC	0.048829	SERP	0.050692	SIGIR	0.031072	PAKDD	0.036814
SIGMOD	0.048045	SIGSOFT	0.048724	SPIRE	0.024243	ICDM	0.031991
DASFAA	0.04765	APSEC	0.047864	CIKM	0.023549	KDD	0.028635
BNCOD	0.04579	CSMR	0.047509	ECDL	0.020604	SSDBM	0.026206
IDEAS	0.044936	ICSE	0.047149	DOCENG	0.018801	SBBD	0.02618

Table 1: An illustration of 4 discovered topics from a 150-topic solution for the corpus. The titles are our interpretation of the topics.

4.3.2 Entropy based Comparison. We provide quantitative comparison between VAT and ACT1 using Entropy. We used average entropy to measure the quality of discovered topics, which reveals the purity of topics, less intra-topic entropy is better.

$$Entropy\ of\ (Topic) = -\sum_z P(z) log_2[P(z)] \quad (5)$$

Fig. 2 shows the average entropy of topic-word distribution for all topics measured by using Eq. 5. Lower entropy for different number of topics T= 100, 150, 200 proves the effectiveness of proposed approach for obtaining better topics. From the curves in Fig. 2 it is clear that VAT outperformed ACT1 for different number of topics. The performance difference for different number of topics is pretty much even, which corroborate that proposed approach's superiority is not sensitive to the number of topics.

4.3.3. Authors Correlations. VAT approach can be used for automatic correlation discovery between authors, including conferences influence in addition to previously used influence of latent topics [16]. Discovered correlations can be utilized to collaborate and work on joint projects.

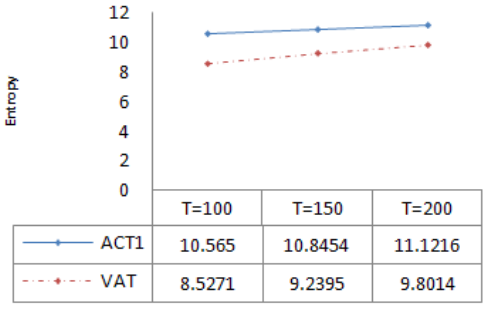

Figure 2: Average Entropy curve as a function of different number of topics, lower is better.

To illustrate how it can be used in this respect, distance between authors i and j is defined as symmetric KL (sKL) divergence between the topics

distribution conditioned on each of the authors distribution as:

$$sKL(i,j) = \sum_{z=1}^{T} \left[\theta_{iz} log \frac{\theta_{iz}}{\theta_{jz}} + \theta_{jz} log \frac{\theta_{jz}}{\theta_{iz}} \right] \quad (6)$$

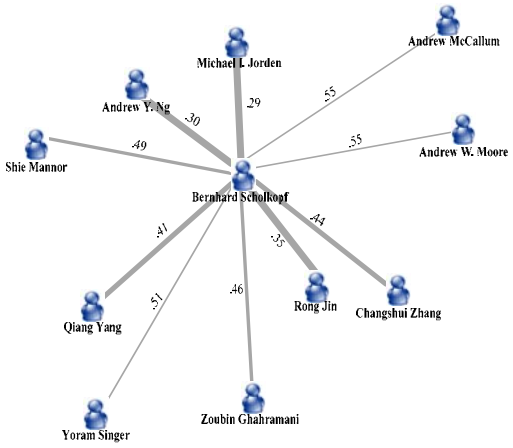

Figure 3: Bernhard Scholkopf correlation graph.

We calculated the dissimilarity between the authors by using Eq. 6 smaller dissimilarity value means higher correlation between the authors. Fig. 3 shows semantic-based correlation graph of Bernhard Scholkopf in which all authors are linked with him through a line showing dissimilarity value, smaller values of linking line shows stronger correlation between the authors. Here, it is obligatory to mention that top 10 authors shown in Bernhard Scholkopf correlation graph are not necessarily the authors who have co-authored with Bernhard Scholkopf mostly, but rather are the authors that tend to produce most words for Bayesian learning topic in the corpus.

It is quite obvious, that top 10 associated authors with Bernhard Scholkopf are well known experts of the machine learning field. In a similar was conferences distribution can be used to find the correlations between conferences.

5. Related Work

Community mining has been a hot issue in social network analysis especially in researcher's social network. On the basis of discovered communities one can solve different kind of academic recommendation tasks. Communities are modeled as graphs and related groups of entities were discovered either by network linkage information [14] or by iterative removal of edges between graphs [10,15,18].

Collaborative filtering [4,7] is employed to discover related groups of entities. They recommended items to the users on the basis of similarity between users and items. Content-based filtering [3] can also be applied to recommend items on the basis of correlations between the content of the items and the user's preferences. This method creates a profile for each item or user to characterize their nature.

Random walk based pair-wise learning [21] and tripartite graph [20] approaches were proposed to discover hidden communities. Recently, community mining problem is investigated including community discovery and change-point detection on dynamic weighted directed graphs [8]. A MetaFac (MetaGraph Factorization), a framework for discovering community structures from social network interactions based on relational hyper graph factorization was proposed [9]. Yang et al., combined link and content analysis for community detection in paper citations and Word Wide Web [19].

Topic-based retrieval approach [13] is applied to match reviewers with the papers without considering conferences information. The importance of conference information is argued by [17] and topic model based approach is proposed. They discovered academics social network by using documents text information while viewing conference information just as a stamp. Aforementioned approaches were incapable of considering implicit semantic information based correlations between text and authors of the conferences; however our proposed CL approach can benefit from it.

Acknowledgements. The work is supported by the Higher Education Commission (HEC), Islamabad, Pakistan.

6. Conclusions

This study deals with the problem of research community mining through semantic analysis of text and relationships between authors from CL. We introduced a topic modeling based VAT approach, which can automatically extract semantically related conferences, authors and topics from the research community. Its generative process links conferences and authors on the basis of latent topics which are quite precise and matches with the real world data. We demonstrated that discovering of authors and topics for new conferences and to finding semantic-based correlations between the authors are functional.

As a future work, we plan to investigate how to add explicit links information on the basis of papers citations, in addition to already used semantic-based information for mining research community.

7. References

1. Andrieu, C., Freitas, N. D., Doucet, A., Jordan, M. An Introduction to MCMC for Machine Learning. JMLR, vol. 50, pp. 5–43, 2003.
2. Azzopardi, L., Girolami, M., Risjbergen, K. van. Investigating the Relationship between Language Model Perplexity and IR Precision-Recall Measures. In Proceedings of the 26th ACM SIGIR, 2003.
3. Balabanovic, M., Shoham, Y. Content-Based Collaborative Recommendation. Communications of the ACM (CACM), vol. 40(3), 1997.
4. Breese, J., Heckerman, D., Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In Proceedings of UA), pp. 43-52, 1998.
5. Blei, D. M., Ng, A. Y., Jordan, M. I. Latent Dirichlet Allocation. JMLR, vol. 3, pp. 993-1022, 2003.
6. DBLP Bibliography database. http://www.informatik.uni-trier.de/~ley/db/.
7. Deshpande, M., Karypis, G.: Item-based Top-n Recommendation Algorithms. ACM Transactions on Information Systems, vol. 22(1), pp. 143-177, 2004.
8. Duan, D., Li, Y., Jin, Y., and Lu, Z. Community Mining on Dynamic Weighted Directed Graphs. In CNIKM Workshop, November 6, 2009.
9. Lin, Y. R., Sun, J., Castro, P., Konuru, R., Sundaram, H., and Kelliher, A. MetaFac: Community Discovery via Relational Hypergraph Factorization. In Proceedings of ACM SIGKDD, June 28– July 1, 2009.
10. Girvan M., Newman, M. E. J. Community Structure in Social and Biological Networks. In Proceedings of the NAS, USA, vol. 99, pp. 8271-8276, 2002.
11. Griffiths, T. L., Steyvers, M. Finding Scientific Topics. In Proceedings of NAS, USA, pp. 5228-5235, 2004.
12. Hofmann, T. Probabilistic Latent Semantic Analysis. In Proceedings of the 15th UAI, Sweden, 1999.
13. Mimno, D., McCallum, A. Expertise Modeling for Matching Papers with reviewers. In Proceedings of the 13th ACM SIGKDD, pp. 500-509, 2007.
14. Pothen, A., Simon, H., Liou, K. P. Partitioning Sparse Matrices with Eigenvectors of Graphs. SIAM Journal on Matrix Analysis and Applications, vol. 11, pp. 430-452, 1990.
15. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D. Dening and Identifying Communities in Networks. In Proceedings of NAS, USA, 2004.
16. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P. The Author-Topic Model for Authors and Documents. In Proceedings of the 20th UAI, 2004.
17. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the 14th ACM SIGKDD, 2008.
18. Tyler, J. R., Wilkinson, D. M., Huberman, B. A. Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. In Proceedings of the International Conference on Communities and Technologies, pp. 81-96, 2003.
19. Yang, T., Jin, R., Chi, Y., and Zhu, S. Combining Link and Content for Community Detection: A Discriminative Approach. In Proceedings of ACM SIGKDD, 2009.
20. Zaiane, O. R., Chen, J., Goebel, R. DBconnect: Mining Research Community on DBLP Data. In Joint 9th WEBKDD and 1st SNA-KDD Workshop, San Jose, California, USA, August 12, 2007.
21. Zhang, J., Tang, J., Liang, B., Yang, Z., Wang, S., Zuo, J., Li, J. Recommendation over a Heterogeneous Social Network. In Proceedings of the 9th International Conference on Web-Age Information Management (WAIM), July 20-22, 2008.