Research Community Mining via Generalized Topic Modeling

ALI DAUD*, MUHAMMAD AKRAM SHAIKH, AND FAQIR MUHAMMAD*****

RECEIVED ON 19.05.2010 ACCEPTED ON 03.01.2011

ABSTRACT

Mining research community on the basis of hidden relationships present between its entities is important from academic recommendation point of view. Previous approaches discovered research community by using network connectivity based distance measures (no text semantics) or by using poorer text semantics and relationships of documents DL (Document Level) by ignoring richer text semantics and relationships of VL (Venue Level). In this paper, we address this problem by considering richer text semantics and relationships. We propose a VAT (Venue Author Topic Approach) based on Author-Topic model to discover inherent community structures in a more realistic way by modeling from VL. We show how topics and authors can be inferred for new venues and how author-to-author and venue-to-venue correlations can be discovered. The positive relationship of topic denseness with ranking performance of proposed approach is explained. Experimental results on research collaborative network "DBLP" demonstrate that proposed approach significantly outperformed the baseline approach in discovering community structures and relationships in large-scale network.

Key Words: **Richer Text Semantics and Relationships, Digital Libraries, Community Mining, Unsupervised Learning**

1. INTRODUCTION

Complex networks exist in diverse domains, such as communication networks, protein interaction networks and social networks. These networks are often comprised of loose clusters (communities), whose members are more strongly connected to each other than the rest of the network. Discovery and identification of these communities is referred to as community mining. Community mining in heterogeneous academic social networks is important problem discussed nowadays, where most of the information is implicit within the entities and their relationships. For example, authors are connected to each other by co-authorships or paper citations and thus can be modeled as interaction graphs. From generic point of view, various conferences are held every year about different topics and huge volume of scientific literature is collected about them in digital libraries which contain hidden community structures. It provides us with many challenging discovery tasks which are very useful from academic recommendation perspective. For example, to find reviewers for a specific venue, suggesting venues to the researchers for submitting papers, inviting program committee members for a conference or suggesting authoritative venues of specific research area to a new researcher for literature reviewing.

* Ph.D. Scholar, Department of Computer Science & Technology, Lab 1-308, FIT Building, Tsinghua University, Beijing, China.
** Professor, Department of Software Engineering, Mehran University of Engineering & Technology, Jamshoro.
*** Professor, Department of Mathematics & Statistics, Allama Iqbal Open University, Sector H-8, Islamabad.

The concept of community is self-explanatory as there is no exact definition that is accepted worldwide. Formerly, two major frameworks used for research community mining (1) investigated the problem by using distance based measures which considers network connectivity on the basis of co-authorship and publishing in the same venue [1-2] and (2) by using latent topics (semantically related probabilistic cluster of words) based models [3-5], without considering venues information. Recently, Tang, et. al. [6] argued that venues and authors are interdependent and should be modeled together. Consequently, a unified topic modeling approach ACT1 (Author Conference Topic) model was proposed, which can discover research community on the basis of semantics-based intrinsic structure of the words and authors by considering venues information. However, they viewed venues information just as a stamp (token), which became the reason of ignoring rich semantics-based text information and authors relationships present between the venues. We think venues implicit semantics-based text information and authors relationships are very important from research community discovery point of view.

In this paper, we investigate this problem by modeling venues richer text semantics and relationships. We generalized previous topic modeling approach [6] from a single document "constituent-document" *(poorer text semantics and relationships)* to all publications of the venue "super-document" *(richer text semantics and relationships)*. The intuition behind considering conferences as super-documents is explained with the help of an example in Fig. 1. Here from text of document means only the title of the paper (instead of using whole paper or abstract) which is usually real representative of the document and contains most important words to explain the main theme of paper. Some preliminary/practical experiments show that there is no significant performance difference if one uses only title words. On the other hand time complexity for model learning is significantly decreased. A constituent-document usually has few semantically related words (as total words are only "8") and authors (as total authors are only "2") to a topic shown in figure 1, while in a conference usually there are many related papers to a topic; as a result super-document usually has many semantically related words (as total words are high in number "439") and authors (as total authors are high in number "95") to a topic shown in Fig. 1. Constituent-document is a subset of super-document as highlighted in Fig. 1; consequently text semantics and relationships are richer in a super-document as compared to a constituent-document.

We propose VAT by using Author-Topic model [7], in which communities are modeled as latent variables and are considered probability distribution on the entire social entity (authors and venues) space, simultaneously. It can be used to discover research communities, make predictions of topics and authors for new venues and can be used to find relationships between authors and relationships between venues. We empirically showed that proposed approach clearly achieve better results than baseline approach due to less sparse topics and solution provided by us produced quite intuitive and functional results. Notably this approach is evaluated in research collaborative network; it can easily be extended to other complex networks-based applications.

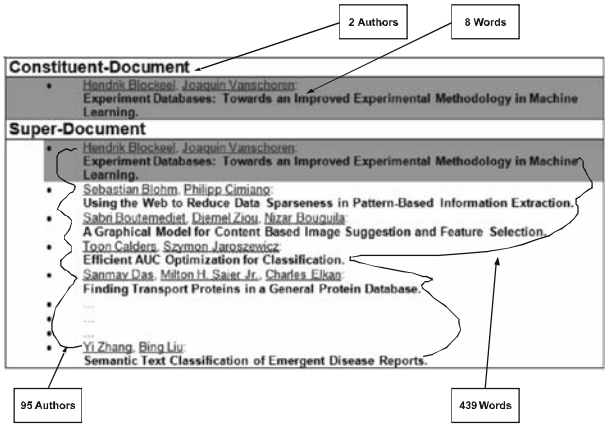


FIG. 1. AN ILLUSTRATION OF RICH TEXT SEMANTICS AND RELATIONSHIPS (ACCEPTED PAPERS BY PKDD-2007)

The novelty of work described in this paper lies in the formalization of the research community mining problem from VL, generalization of previous topic modeling approach from DL to VL (VAT) for capturing richer text semantics and relationships, and experimental verification of the effectiveness of 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, which can capture word-to-word, word-to-author, word-to-venue, author-to-author and author-to-venue correlations.

The rest of the paper is organized as follows. In Section 2, we formalize research community mining problem. Section 3 provides background and illustrates proposed approach for modeling research community with its parameter estimation and inference making details. In Section 4, corpus, parameters settings, performance measures with empirical studies and discussions about the results are given. Section 5 provides related work and Section 6 brings this paper to the conclusions and provides future work.

Note that in the rest of the paper, we use the term constituent-document, accepted paper, publication and document interchangeably. Here Venue can be a conference or journal (our focus is on conferences here). Additionally "super-document" means all the documents of one conference.

2. PROBLEM SETTING

Community mining is becoming more and more interesting with the emergence of various digital scientific libraries. Our work is focused on mining research community by modeling relationships between its entities on the basis of implicit semantics-based text information and relationships present between the venues. Each venue accepts many papers every year written by different authors. To our interest, each publication contains title and authors names. Venues with their accepted papers on the basis of latent topics can help us to discover communities. Fig. 2 graphically shows how authors and conferences can build up communities on the basis of latent topics, where one community consists of topically related authors and conferences.

We denote a venue c as a vector of N_c words based on the accepted papers, an author of a paper as r, and formulize the problem as: Given a venue c with N_c words, and $\mathbf{a}_c$ authors, discover research community. Formally for finding topically related venues and authors, we need to calculate the probability $p(z|c)$, $p(z|r)$ and $p(w|z)$ where z is a latent topic, r is an author and w is the words of super-document.

3. LATENT TOPIC BASED COMMUNITY MODELING

In this section, before describing proposed approach, we will first describe how documents, authors and venues are modeled with the help of latent topics.

3.1. Latent Dirichlet Allocation

Fundamental topic modeling approach LDA (Latent Dirichlet Allocation) [8] assumes that there is a hidden topic layer $Z=\{z_1, z_2, z_3, \ldots, z_t\}$ between the word tokens

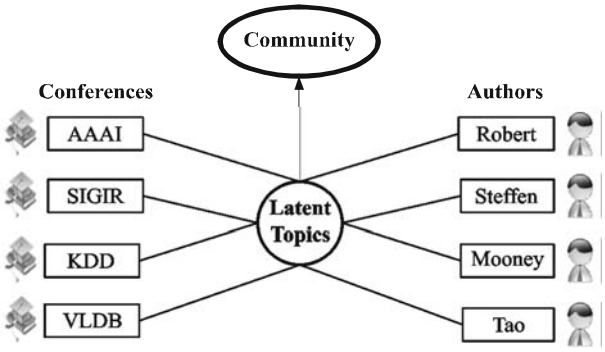


FIG. 2. RESEARCH COMMUNITY MINING

and documents, where z_i denotes a latent topic and each document d is a vector of N_d words w_d. A collection of D documents is defined by $D=\{w_1, w_2, w_3,\ldots, w_d\}$ and each word w_{id} is chosen from a vocabulary of size V. First, for each document d, a multinomial distribution θ_d over topics is randomly sampled from a Dirichlet distribution with parameter α. Second, for each word w, a topic z is chosen from this topic distribution. Finally, the word w is generated by randomly sampling from a topic-specific multinomial distribution Φ_z. The generating probability of word w from document D for LDA is given as:

$$P\left(w\middle|d,\theta,\phi\right) = \sum_{z=1}^{T} P\left(w\middle|z,\phi_z\right)P\left(z\middle|d,\theta_d\right) \quad (1)$$

3.2. Author Topic Model

Following topic modeling basic idea of modeling words and documents, words and authors are modeled by considering latent topics to discover the research interests of authors [7]. In AT (Author Topic) model, each author (from set of A authors) of a document d is associated with a multinomial distribution θ_a over topics is sampled from Dirichlet α and each topic is associated with a multinomial distribution Φ_z sampled from Dirichlet β over words of a document for that topic. The generating probability of word w for author r of a document d is given in Equation (2). It has successfully discovered topically related authors but did not consider venues information.

$$P\left(w\middle|r,d,\phi,\theta\right) = \sum_{z=1}^{T} P\left(w\middle|z,\Phi_z\right)P\left(z\middle|r,\theta_r\right) \quad (2)$$

3.3 Author Conference Topic Model

Consequently, ACT1 was proposed to model authors and venues (communities in research collaborative network) together [6]. In this model, each author is represented by the probability distribution θ_d over topics and each topic is represented as a probability distribution Φ_z over words and ψ_z over conferences for each word of a document for that topic. The generating probability of word w and conference c for author r of a document d is given in Equation (3). Here, venue is viewed as a stamp associated with each word with same value. So, the modeling is just based on semantics-based text information and co-authorship of documents, while rich semantics-based structure of words and authors correlations present between venues on the basis of publishing in the same conference was ignored, which motivated us to purpose VAT.

$$P\left(w,c\middle|r,d,\phi,\psi,\theta\right)=\sum_{z=1}^{T} P\left(w\middle|z,\Phi_z\right)P\left(c\middle|z,\psi_z\right)P\left(z\middle|r,\theta_r\right) \quad (3)$$

3.2 Venue Author Topic Approach

By using hidden topic layer one can capture the semantic information present in the text to model multiple entities at once. The basic idea presented in AT model [7], that words and authors of documents can be modeled by considering latent topics became the intuition of modeling words, authors and venues, simultaneously. In the proposed approach, we viewed a venue as a composition of its all documents words and authors of its publications. Symbolically, for a venue C (a super-document) we can write it as: $C=\{(\mathbf{w}_1,\mathbf{a}_{d1})+(\mathbf{w}_2,\mathbf{a}_{d2})+(\mathbf{w}_3,\mathbf{a}_{d3})+\ldots+(\mathbf{w}_i,\mathbf{a}_{di})\}$, where w_i is a word vector of document for a venue and a_{di} are author(s) of that document.

DL approach considers that an author is responsible for generating some latent topics of the documents on the basis of semantics-based information present in the text and co-authorship based correlations. While, VL approach considers that an author is responsible for generating some latent topics of the venue on the basis of rich semantics-based information present in the text as well as rich co-venue based correlations (Fig. 3(a-b)). In VAT, each author (from set of K authors) of a venue is associated with a multinomial distribution θ_r over topics and each topic is associated with a multinomial distribution Φ_z over words

of a venue for that topic. Both θ_r and Φ_z have symmetric Dirichlet prior with hyper parameters α and β. The generating probability of the word *w* for author *r* of a venue c is given as:

$$P(w|r,c,\phi,\theta) = \Sigma_{z=1}^{T} P(w|z,\Phi_z) P(z|r,\theta_r) \quad (4)$$

The generative process of VAT is as follows:

For each author $r=1,\ldots, K$ of venue *c*

Choose θ_r from Dirichlet (α)

For each topic $z=1,\ldots,T$

Choose Φ_z from Dirichlet (β)

For each word $w=1,\ldots, N_c$ of venue *c*

Choose an author r uniformly from all authors $\mathbf{a}_c$

Choose a topic *z* from multinomial (θ_r) conditioned on *r*

Choose a word w from multinomial (Φ_z) conditioned on *z*

Gibbs sampling is utilized [9,7] 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:

$$P(z_i=j, r_i=k | w_i=m, \mathbf{z}_{-i}, \mathbf{r}_{-i}, \mathbf{a}_c) \propto \frac{n_{-i,j}^{(wi)} + \beta n_{-i,j}^{(ri)} + \alpha}{n_{-i,j}^{(.)} + w\beta n_{-i,.}^{(ri)} + R\alpha} \quad (5)$$

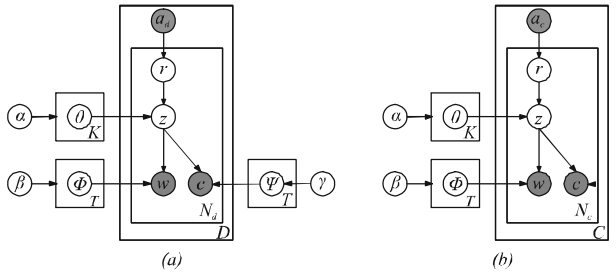


FIG. 3(a). ACT1 (DL COMMUNITY MODELING) (b) VAT (VL COMMUNITY MODELING)

where $z_i=j$ and $r_i=k$ represent the assignments of the i^{th} word in a venue to a topic j and author k respectively, $w_i=m$ represents the observation that ith word is the m^{th} word in the lexicon, and z_{-i} and r_{-i} represents all topic and author assignments not including the ith 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, *W* is the size of the lexicon and *R* is the number of authors. "." Indicates summing over the column where it occurs and $n_{-i,j}^{(.)}$ stands for number of all words that are assigned to topic *z* excluding the current instance.

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

$$\Phi_{zw} = \frac{n_{-i,j}^{(wi)} + \beta}{n_{-i,j}^{(.)} + w\beta} \quad (6)$$

$$\theta_{rz} = \frac{n_{-i,j}^{(ri)} + \alpha}{n_{-i,.}^{(ri)} + R\alpha} \quad (7)$$

where, ϕ_{zw} is the probability of word w in topic z and θ_{rz} is the probability of topic *z* for author *r*. These values correspond to the predictive distributions over new words w and new topics z conditioned on *w* and *z*. To find *Z*x*C* (topic by venue) count matrix we calculated the distribution of topic given venue as:

$$p(z|c) = \Sigma_{r \in R_c} p(z|r) p(r|c) = \frac{1}{|R_c|} \Sigma_{r \in R_c} p(z|r) \quad (8)$$

where r_c is the number of authors belongs to a venue *c*.

4. EXPERIMENTS

4.1 Corpus

We downloaded five years publication corpus of conferences from DBLP [10ke]. In total, we extracted 112,317 authors, 90,124 publications, and combined them into a super-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. This led to a vocabulary size of V=10,872, a total of 572,592 words and 26,078 authors in the corpus. Fig. 4 shows fairly smooth yearly data distribution for number of publications (D) and authors (R) in conferences.

4.2 Parameter Settings

The best possible values of hyper-parameters α and β (Fig. 3(b)) can be estimated by using Expectation-Maximization [11] or Gibbs sampling algorithm [9]. Expectation-Maximization algorithm is susceptible to local maxima and computationally inefficient [8], consequently we use Gibbs sampling algorithm. In our Gibbs sampling algorithm based experiments, for 150 topics Z the hyper-parameters α and β were set at $50/Z$ and 0.1 respectively, by following the values used in [7]. The number of topics Z was fixed at 150 on the basis of human judgment of meaningful topics and measured perplexity [12].

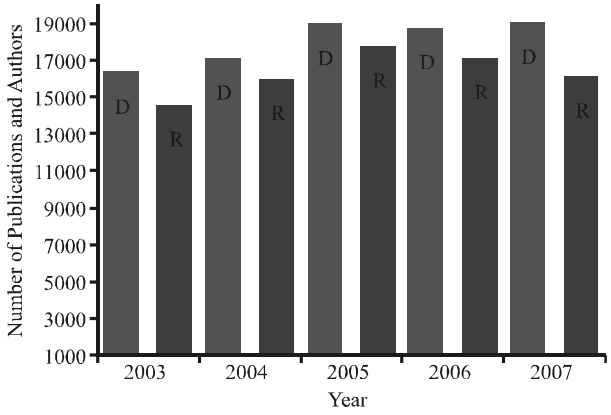


FIG. 4. HISTOGRAM ILLUSTRATING DATA DISTRIBUTION

4.3 Performance Measures

Perplexity is usually used to measure the performance of latent topic based models; however, it cannot be a statistically significant measure when they are used for information retrieval [12]. So we used Entropy, sKL Divergence and Error Rate to measure the performance. In our experiments, at first we used average entropy to measure the quality of discovered topics, which reveals the purity of topics, less intra-topic entropy is better. Secondly, we used average Symmetric KL (sKL) divergence [7,13] to measure the quality of topics, in terms of inter-topic distance, more inter-topic sKL divergence (distance) is better.

To measure the performance in terms of precision and recall [12] is out of question due to unavailability of standard dataset and use of human judgments cannot provide appropriate (unbiased) answers for performance evaluation. Consequently, we used a simple error rate method to evaluate the performance in terms of authors and conferences ranking. We discovered top 7 authors and conferences related to top most author (e.g. for VAT XMLDB topic it is Wei Wang) and top most conference (e.g. for VAT XMLDB topic it is Xsym) in each topic by using sKL divergence (Table 1). We compared these top 7 authors and conferences with topically discovered top 8 authors and conferences and calculated error rate with respect to their absence or presence in the topically ranked authors and conferences list in Table 1.

$$\text{Entropy of Topic } = -\Sigma_z P(z) \log_2 \left[P(z) \right] \tag{9}$$

$$\text{sKL}(i, j) = \Sigma_{z=1}^{T} \left[\theta_{iz} \log \frac{\theta_{iz}}{\theta_{jz}} \log \frac{\theta_{jz}}{\theta_{iz}} \right] \tag{10}$$

TABLE 1. AN ILLUSTRATION OF 5 DISCOVERED TOPICS (TOP VAT, BOTTOM ACT1) FROM A 150-TOPIC SOLUTION FOR THE CORPUS. EACH TOPIC IS SHOWN WITH THE TOP 8 WORDS, AUTHORS AND VENUES THAT HAVE HIGHEST PROBABILITY CONDITIONED ON THAT TOPIC. THE TITLES ARE OUR INTERPRETATION OF THE TOPICS. HERE ACRONYMS ARE XMLDB (XML DATABASES), SE (SOFTWARE ENGINEERING), DM (DATA MINING), BL (BAYESIAN LEARNING) AND WS (WEB SEARCH)

Topic 54 (VAT)" XML Databases"		Topic 100 (VAT)" Software Engineering"		Topic 99 (VAT)" Data Mining"		Topic 102 (VAT)" Bayesian Learning"		Topic 15 (VAT)" Web Search"	
Word	Prob.	Word	Word	Word	Prob.	Word	Prob.	Word	Prob.
Xml	0.093811	Software	0.197374	Mining	0.15788	Learning	0.215469	Web	0.56497
Data	0.079758	Development	0.059311	Clustering	0.10328	Bayesian	0.044161	Search	0.02853
Query	0.072938	Engineering	0.053288	Data	0.070152	Classification	0.033074	Social	0.021536
Queries	0.053479	Component	0.035498	Classification	0.049702	Kernel	0.023655	Engine	0.015482
Databases	0.051068	Testing	0.032255	Patterns	0.037126	Markov	0.017572	Collaborative	0.01212
Database	0.042586	Agile	0.027159	Frequent	0.028128	Feature	0.014727	Pages	0.01212
Processing	0.026536	Test	0.026695	Discovery	0.027515	Supervised	0.014531	Personalized	0.011851
Relational	0.025205	Requirements	0.026047	Association	0.021993	Clustering	0.01404	Information	0.00916
Author	Prob.	Author	Prob.	Author	Prob.	Author	Prob.	Author	Prob.
Wei Wang	0.011326	Frank Maurer	0.013063	Philip S. Yu	0.021991	Bernhard Scholkopf	0.014908	Katsumi Tanaka	0.012123
Divesh Srivastava	0.010205	Mario Piattini	0.009204	Reda Alhajj	0.01785	Michael I. Jordan	0.011987	Wolfgang Nejdl	0.012123
Elke A. Rundensteiner	0.009747	Baowen Xu	0.007954	Jiawei Han	0.017792	Rong Jin	0.010863	Qing Li	0.011791
Kian-Lee Tan	0.008881	Steacute Ducasse	0.006976	Hans-Peter Kriegel	0.014686	Andrew Y. Ng	0.009739	Amit P. Sheth	0.011259
Sourav S. Bhowmick	0.008779	John C. Grundy	0.006812	Wei Wang	0.01066	Zoubin Ghahramani	0.008222	Boi Faltings	0.011259
Divyakant Agrawal	0.008269	Grigori Melnik	0.006323	Eamonn J. Keogh	0.010373	Qiang Yang	0.007773	C. Lee Giles	0.009996
Nick Koudas	0.008218	Gerardo Canfora	0.006269	Christos Faloutso	0.009567	Sebastian Thrun	0.00693	Ning Zhong	0.009796
Gerhard Weikum	0.007913	Arie van Deursen	0.005399	Ming-Syan Chen	0.008992	Yoram Singer	0.006649	Marco Brambilla	0.008134
Conference	Prob.	Conference	Prob.	Conference	Prob.	Conference	Prob.	Conference	Prob.
Xsym	0.094329	Agile Deve.	0.104955	SDM	0.043478	ALT	0.056253	LA-WEB	0.029552
SSDBM	0.06592	XP	0.076159	DAWAK	0.042625	UAI	0.053589	WISE	0.026334
ADBIS	0.050305	WCRE	0.058275	PKDD	0.03843	COLT	0.051084	WIDM	0.025327
ADC	0.048829	SERP	0.050692	PAKDD	0.036814	NIPS	0.049047	WI	0.023486
SIGMOD	0.048045	SIGSOFT	0.048724	ICDM	0.031991	ICML	0.047252	ISWC	0.023061
DASFAA	0.04765	APSEC	0.047864	KDD	0.028635	ECML	0.044731	ASWC	0.023061
BNCOD	0.04579	CSMR	0.047509	SSDBM	0.026206	PKDD	0.025322	WWW	0.020801
IDEAS	0.044936	ICSE	0.047149	SBBD	0.02618	SDM	0.024587	ICWS	0.018834
Topic 4 (ACT1)" XML Databases"		Topic 71 (ACT1)" Software Engineering"		Topic 7 (ACT1)" Data Mining"		Topic 62 (ACT1)" Bayesian Learning"		Topic 130 (ACT1)" Web Search"	
Word	Prob.	Word	Word	Word	Prob.	Word	Prob.	Word	Prob.
data	0.031350	Agile	0.028619	Data	0.030168	Learning	0.049237	Web	0.059021
xml	0.031176	Software	0.023352	Mining	0.024621	Data	0.009555	Based	0.015124
query	0.023387	Development	0.018336	Clustering	0.021486	Based	0.009333	Search	0.014413
database	0.018020	Based	0.016329	Patterns	0.008706	Classification	0.008890	Semantic	0.012281
web	0.013000	Component	0.014573	Classification	0.007982	Models	0.008890	Xml	0.006949
system	0.012135	Programming	0.010309	Based	0.007982	Kernel	0.008890	Hypermedia	0.006238
processing	0.011789	Extreme	0.008302	Frequent	0.007259	Clustering	0.008003	Information	0.006238
based	0.011096	Systems	0.008052	Learning	0.006053	Bayesian	0.007116	Services	0.005883
Author	Prob.	Author	Prob.	Author	Prob.	Author	Prob.	Author	Prob.
Surajit Chaudhuri	0.010518	Frank Maurer	0.009771	Philip S. Yu	0.010914	Shie Mannor	0.004564	Frank M. Shipman	0.004331
Anastassia Ailamaki	0.006414	Pekka Abrahamsson	0.007271	Vipin Kumar	0.007087	James T. Kwok	0.003904	Zheng Chen	0.004029
Kevin Chen-Chuan	0.006347	Mike Holcombe	0.005374	Pang-Ning Tan	0.004961	Satinder P. Singh	0.003904	Wendy Hall	0.003802
Elke A. Rundensteiner	0.005876	Yael Dubinsky	0.004771	George Karypis	0.004876	Sridhar Mahadevan	0.003822	Amit P. Sheth	0.003424
Raghu Ramakrishnan	0.005270	Stefan Biffl	0.004426	Martin Ester	0.004706	Lawrence Carin	0.003574	C. Lee Giles	0.003424
Wenfei Fan	0.004799	Richard F. Paige	0.003132	Eamonn J. Keogh	0.004706	Andrew Y. Ng	0.003492	Irwin King	0.003047
H. V. Jagadish	0.004463	James Miller	0.003132	Wei Fan	0.004450	Michael I. Jordan	0.003409	Ji-Rong Wen	0.002895
Jeffrey F. Naughton	0.004328	Rick Mugridge	0.003132	Srinivasan	0.004450	Bernhard Scholkopf	0.003327	Altigran Soares	0.002744
Conference	Prob.	Conference	Prob.	Conference	Prob.	Conference	Prob.	Conference	Prob.
VLDB	0.375615	EUROMICRO	0.517740	SDM	0.618306	ICML	0.980336	WWW	0.729783
SIGMOD	0.305755	XP	0.331652	ICDE	0.235965	NIPS	0.004367	HYPERTEXT	0.199405
ICDE	0.296397	AGILE	0.130568	KDD	0.123170	KDD	0.002632	WISE	0.064734
XSYM	0.008406	ICSE	0.008214	VLDB	0.010051	AAAI	0.002342	SACMAT	0.000240
SBBD	0.006147	CAISE	0.001397	HIPC	0.002295	COLT	0.001475	CIKM	0.000240
WWW	0.002114	CATA	0.000375	ICML	0.000679	UAI	0.000896	SIGIR	0.000022
KDD	0.0005	CIC	0.000375	ICAIP	0.000356	CIC	0.000318	CAISE	0.000022
ADBIS	0.000339	AOSD	0.000375	APSEC	0.000356	OOPLSA	0.000318	KI	0.000022

4.4 Baseline Approach

We compared proposed VAT with ACT1 and used same number of topics for comparability. The number of Gibbs sampler iterations used for ACT1 is 1000 and parameter values same as the values used in [6].

4.5 Results and Discussion

4.5.1 Mined Community

The effect of rich text semantics and relationships on the performance of topic modeling approach is studied both qualitatively and quantitatively for community mining problem. Firstly, we provide qualitative comparison between VAT and ACT1. We extracted and probabilistically ranked authors and venues related to specific area of research on the basis of latent topics. Table 1 illustrates 5 different topics out of 150, discovered from the 120th iteration of particular Gibbs sampler run. Here it is necessary to mention that usually in DL topic modeling 1000 [14] and 2000 [6,7] number of Gibbs sampling iterations are used for 28154, 160000 and 10716 documents respectively, as they are large number of documents. While we have only 261 super-documents which are with wealthier text semantics and relationships, as a result we are able to obtain fine grained topics after much smaller number of iterations.

The words associated with each topic for VAT are strongly semantically related (less sparse) than that of ACT1, as they are assigned higher probabilities (Table 1). Illustratively, words associated with "Web Search" topic discovered by VAT is very much clear about searching information on the web, while "Web Search" topic discovered by ACT1 is not clear as web search, web services and XML are mixed in one topic. ACT1 faces the same problem of topic sparseness for other discovered topics encompassed in the corpus (Fig. 4 to see quantitative comparison of topic sparseness).

Additionally, it is noted that because of topic sparseness topically related authors and conferences (communities) are also sparse (not from the specific area of research). Consequently, the authors and conferences associated with the topics for VAT are more intuitive than ACT1. For example, VAT discovers Philip S. Yu, Jiawei han and Reda Alhajj for "Data Mining" topic because it has only one topic for data mining, while ACT1 cannot, because ACT1 discovered six different topics of data mining (based on top ten words) with almost similar kind of words but different associated authors, that became the reason of sparseness of authors and Jiawei Han and Reda Alhajj are not even assigned to same data mining topic in the remaining data mining discovered by ACT1. For conferences, top eight related conferences to "Data Mining" topic of VAT are highly related, except last two which are more related to databases (SSDBM, SBBD), while top eight conferences related to "Data Mining" topic of ACT1 includes only top three conferences related to data mining and others are related to databases (VLDB), HIPC (High Performance Computing), artificial intelligence (ICML, ICAIP) and software engineering (APSEC). Similarly, for "Web Search" topic VAT discovers almost all related conferences, while ACT1 model only discovers top three related conferences to web search and others are related to access control methods (SACMAT), information retrieval (CIKM, SIGIR), software engineering (CAISE) and artificial intelligence (KI). These results shows that the denseness of topics is a function of discovering compact communities, conversely sparseness of topics is a function of discovering less compact communities, which concludes the more sparse the topics the poor the approach will perform.

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 venues that are semantically related to the topic, which build up a topic based community.

In addition to qualitative comparison, we also provide quantitative comparison between VAT and ACT1. Fig. 5(a) shows the average entropy of topic-word distribution for all topics measured by using Equation (6). Lower entropy for different number of topics T=100, 150, 200 proves the effectiveness of proposed approach for obtaining better topics. Fig. 5(b) shows the average distance of topic-word distribution between all pairs of the topics measured by using Equation (7). Higher sKL divergence for different number of topics T=100, 150, 200 confirms the effectiveness of proposed approach for obtaining compact topics, which resulted in its better ranking performance shown in Table 1.

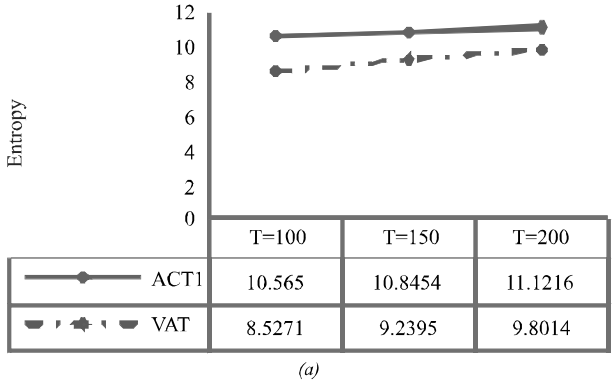
From the curves in Fig. 5(a-b) it is clear that VAT outperformed ACT1 for different number of topics. The performance difference for different number of topics is pretty much even, which corroborate that proposed approach's superiority is not sensitive to the number of topics.

Fig. 5(a) Average Entropy curve as a function of different number of topics, lower is better and Fig. 5(b) Average sKL divergence curve as a function of different number of topics, higher is better.

4.5.2 Topics and Authors for New Venues

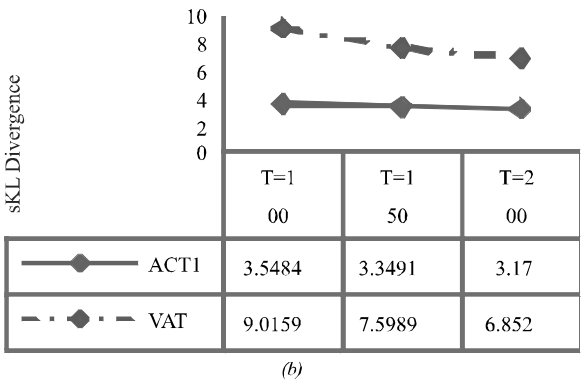
One would like to quickly access the topics and authors for new venues not contained in the training corpus. For this purpose we apply Equation (5) only on the word tokens and authors in new venue each time temporarily updating the count matrices of (word by topic) and (topic by author). The resulting assignments of words to topics can be saved after a few iterations (10 in our simulations). Then Equation (8) is used to calculate the count matrix of (topic by venue). Table 2 shows this type of inference. To show predictive power of VAT we treated two venues as test venues one at a time, by training model on remaining 260 venues.

Predicted words associated with each topic and authors are quite intuitive, as they provide a summary of a specific area of research and are true representatives of the venues. For example, AAAI is one of the best conferences in the area of Artificial Intelligence. Top two predicted topics and their related authors are very insightful, as "Bayesian Learning" has been a main focus of artificial intelligence these years and Bernhard Scholkopf, Michael I. Jordan etc are doubtlessly well-known persons of this area of research. Second topic for this conference is "Semantic Web" which shows the that Bayesian learning is used in abundance to model semantic web and Yong Yu, Katsumi



	T=100	T=150	T=200
ACT1	10.565	10.8454	11.1216
VAT	8.5271	9.2395	9.8014

(a)



	T=100	T=150	T=200
ACT1	3.5484	3.3491	3.17
VAT	9.0159	7.5989	6.852

(b)

FIG. 5(a). AVERAGE ENTROPY CURVE AS A FUNCTION OF DIFFERENT NUMBER OF TOPICS, LOWER IS BETTER AND (b) AVERAGE SKL DIVERGENCE CURVE AS A FUNCTION OF DIFFERENT NUMBER OF TOPICS, HIGHER IS BETTER

Tanaka etc are the authors who produced most words for the semantic web topic or are very much active in this area of research. Topics and authors predicted for VLDB conference are also intuitive and precise, as they are very much representative of the conference sub areas of research in the real-world. Comparatively ACT1 is unable to directly predict topics and authors for new conferences.

4.5.3 Authors and Venues Correlations and Effect of Rich Text Semantics and Relationships

VAT can be used for automatic correlation discovery between authors and venues by including conferences influence in addition to previously used influence of latent topics [7]. To illustrate how it can be used in this respect, distance between authors *i* and *j* and venues *i* and *j* is calculated by using Equation (7) for author-topic and venue-topic distributions, respectively.

We provide correlations based comparison by calculating error rate. Tables 3-4 shows top seven authors and conferences related to top author and conference for each topic of VAT and ACT1 by using sKL divergence. For example in case of XMLDB topic Haixun Wang, Jian Pei, Jun Zhang, Jiawei Han, Wee Keong N, Raymond K. Wong and Ada Wai-Chee Fu are top seven authors correlated to Wei Wang and SSDBM, SIGMOD, ADBIS, DASFAA, SBBD, IDEAS, VLDB are top seven conference related to Xsym for VAT.

The highlighted blocks in Tables 3-4 shows that similar results are obtained for discovered topics in Table 1 for sKL divergence calculated for top most author and conference. For example, in case of VAT top 8 authors shown in Table 1 for XMLDB topic has no authors in common, and for SE topic three authors are common, which are Grigori Melnik, Steacute Ducasse and Mario Piattini, so on. From top 7 authors and venues for five selected topics shown in Tables 3-4 the overall ER (Error Rate) for VAT is less (28.03 for authors, 57.14 for conferences) than ACT1. Its shows that correlations discovered from VL are more precise and bad effect of topic sparseness on the baseline approach for discovering correlations.

5. RELATED WORK

5.1 Community Mining

Community mining has been a hot issue in social network analysis. Communities are modeled as graphs and related groups of entities were discovered either by network

TABLE 2. AN ILLUSTRATION OF TOP 2 PREDICTED TOPICS FOR AAAI AND VLDB CONFERENCES; EACH TOPIC IS SHOWN WITH TITLE (OUR INTERPRETATION OF THE TOPIC), TOP SIX RELATED AUTHORS AND TOP 10 WORDS

AAAI
"Bayesian Learning"
learning, bayesian, classification, kernel, markov, clustering, inference, regression, vector, Gaussian Bernhard Scholkopf, Michael I. Jordan, Andrew Y. Ng, Zoubin Ghahramani, Zhi-Hua Zhou, Rong Jin
"Semantic Web"
web, semantic, ontology, ontologies, owl, search, semantics, rdf, pages, social Yong Yu, Katsumi Tanaka, C. Lee Giles, Wolfgang Nejdl, Ning Zhong, Ian Horrocks
VLDB
"XML Databases"
xml, query, data, queries, databases, database, processing, relational, indexing, documents Kian-Lee Tan, Divesh Srivastava, Sourav S. Bhowmick, Elke A. Rundensteiner, Dongqing Yang
"Data Mining"
mining, data, clustering, classification, patterns, frequent, association, discovery, text, rules Wei Wang, Philip S. Yu, Hans-Peter Kriegel, Reda Alhajj, Jiawei Han, Ming-Syan Chen

linkage information [15-18] or by iterative removal of edges between graphs [19-21], where distance-based measures are utilized. Some approaches used centrality indices distance-based measure for finding related communities [22-23].

Collaborative filtering [24-25] is employed to discover related groups of entities. Content-based filtering [26] can also be applied to recommend items on the basis of correlations between the content of the items and the user's preferences.

TABLE 3. AN ILLUSTRATION OF 5 TOPICS SPARSENESS FOR TOPICALLY RELATED AUTHORS DISCOVERY IN TERMS OF ERROR RATE (LOWER IS BETTER)

		VAT Approach		
XMLDB	**SE**	**DM**	**BL**	**W S**
Haixun Wang	Grigori Melnik	Hans-Peter Kriegel	Michael I. Jordan	Ricardo A. Baeza.
Jian Pei	Steacute Ducasse	Kotagiri Ramamoh.	Andrew Y. Ng	Weiyi Meng
Jun Zhang	Mario Piattini	Heikki Mannila	Rong Jin	Ning Zhong
Jiawei Han	Mike Holcombe	Taneli Mielik.	Qiang Yang	Marco Brambilla
Wee Keong N	Giancarlo Succi	Reda Alhajj	Changshui Zhang	Wolfgang Nejdl
Raymond K. Wong	Xavier Franch	Eamonn J. Keogh	Zoubin Ghahramani	Amit P. Sheth
Ada Wai-Chee Fu	Xudong He	Ming-Syan Chen	Shie Mannor	Qing Li
ER=100	ER=57.14	ER=42.85	ER=28.57	ER=42.85
Average Error Rate= 54.82%				
ACT1				
XMLDB	SE	DM	BL	WS
Anastassia Ailamaki	Yael Dubinsky	Peer Kroger	Robert E. Schapire	Kaj Gronbaek
Wenfei Fan	Mike Holcombe	Jianyong Wang	Naftali Tishby	Maria Bielikova
Jayavel Shanmug.	Helen Sharp	Ruoming Jin	Peter L. Bartlett	Erik Wilde
Philip A. Bernstein	Rick Mugridge	Sanjay Chawla	Gilles Blanchard	Wendy Hall
AnHai Doan	Bartosz Walter	Haixun Wang	Yoram Singer	Weigang Wang
Michael J. Carey	Michele Marchesi	Xifeng Yan	Thomas G. Dietterich	Masashi Toyoda
Renee J. Miller	Laurie Williams	Jiawei Han	John Langford	Nikos Karousos
ER=71.42	ER=57.14	ER=100	ER=100	ER=85.71
Average Error Rate= 82.85%				

TABLE 4. AN ILLUSTRATION OF 5 TOPICS SPARSENESS FOR TOPICALLY RELATED CONFERENCES DISCOVERY IN TERMS OF ERROR RATE (LOWER IS BETTER)

VAT Approach					ACT1 Approach				
XMLDB	SE	DM	BL	WS	XMLDB	SE	DM	BL	WS
SSDBM	XP	ICDM	COLT	WWW	ICDE	APSEC	ICDM	ECML	LA-WEB
SIGMOD	ICSE	PAKDD	NIPS	WI	SIGMOD	SERP	PAKDD	NIPS	W
ADBIS	WCRE	KDD	UAI	WIDM	DASFAA	SEKE	KDD	UAI	ICWS
DASFAA	SERP	PKDD	ICML	ISWC	DEXA	ICSOC	PKDD	ALT	ISWC
SBBD	CSMR	DS	ECML	ASWC	IDEAS	AOSD	DS	AAAI	ASWC
IDEAS	APSEC	DAWAK	DS	ICWS	WAIM	ICSE	ECML	DS	SEBD
VLDB	SIGSOFT	ECML	KDD	SAC	BNCOD	CSMR	ICDE	AI	SAC
ER=14.28	ER=0	ER=28.57	ER=28.56	ER=14.28	ER=71.42	ER=71.42	ER=71.42	ER=57.14	ER=100
Average Error Rate=17.14%					Average Error Rate=74.28%				

Recently, random walk based; pair-wise learning [2] and tripartite graph [1] approaches were proposed to discover hidden communities. Community discovery in large social and information networks has been performed by studying the statistical properties [27] and scalable community is discovered by using text data with correlations [28].

Community mining problem is investigated including community discovery and change-point detection on dynamic weighted directed graphs [29.]. A MetaFac (MetaGraph Factorization), a framework for discovering community structures from social network interactions based on relational hyper graph factorization was proposed [30]. Yang, et.al., combined link and content analysis for community detection in paper citations and WWW (Word Wide Web) [31].

Topic modeling based probabilistic approaches [3-5,32] are applied to discover communities successfully without considering venues information. The importance of venues information is argued and a unified topic model ACT1 is proposed [6], which uses conferences information. They discovered academics social network by using documents information while viewed conference information just as a stamp. Previous approaches were incapable of considering implicit semantics-based rich intrinsic structure of words and rich relationships present between text and authors of venues; however proposed approach can benefit from it by directly modeling from VL.

5.2 Topic Modeling

Automatic extraction of topics from text is performed by [17,33] to cluster documents into groups based on similar semantic contents. Clustering provides a good way to group similar documents in one specific cluster, while particularly a document can have more than one topic e.g. this paper at least has two topics; which are community discovery and topic modeling.

For this reason soft clustering technique PLSA (Probabilistic Latent Semantic Analysis) [11] was proposed as a probabilistic alternative to projection and clustering methods. It can assign each document to almost all clusters with higher or lower probabilities by using Expectation Maximization algorithm. It was generative only at words level but not at documents level, so it was not clear to assign a probability to a document outside the corpus and number of parameters in the model grows linearly with the size of corpus.

Consequently, a probabilistic topic model LDA was proposed [8], which was generative at both words and document level and does not has linear parameters growth problem with the input data. Later, LDA was extended to Author-Topic model [7] for modeling the interests of authors on the basis of latent topics; however we used it to discover research community.

6. CONCLUSIONS

This study deals with the problem of research community mining through capturing richer text semantics and relationships. A generalized topic modeling approach VAT by using Author-Topic model is proposed to handle this problem. We conclude that our generalization from DL to VL is innovative, as discovered communities through proposed approach (can also be applied to journals dataset such as HEP or OHSUMED) is better than the baseline. While, predicted authors and topics for new conferences are practical and meaningful. Our approach (capturing VL text semantics and relationships) was also proved effective in finding authors and conferences correlations when compared with the baseline approach (capturing DL text semantics and relationships). We studied the effect of generalization on topics denseness when modeling community and concluded that dense topics will result in better performance of the approach. Empirical results show overall better performance of VAT on the basis of richer

text semantics and relationships as compared to baseline approach. Even though proposed approach is quite simple, nonetheless it reveals interesting information about several academic recommendation tasks.

From generic point of view, our approach can also be applied to blogs dataset for bloggers interests discovery, news dataset for discovering news reporters interests and active news issues and decisively any dataset which has text and composing authors information. As a possible future direction VAT can be extended by adding time information for evolutionary community mining.

ACKNOWLEDGEMENTS

The work is supported by the HEC (Higher Education Commission), Pakistan. Authors are thankful to Jing Zhang and Feng Wang, Tsinghua University, China for valuable discussions and suggestions.

REFERENCES

[1] Zaiane, O.R., Chen, J., and Goebel, R.D., "Bconnect: Mining Research Community on DBLP Data", Joint 9th WEBKDD and 1st SNA-KDD Workshop, San Jose, California, USA, August 12, 2007.

[2] Zhang, J., Tang, J., Liang, B., Yang, Z., Wang, S., Zuo, J., and Li, J., "Recommendation Over a Heterogeneous Social Network", Proceedings of WAIM, China, 2008.

[3] Mimno, D., and McCallum, A., "Expertise Modeling for Matching Papers with Reviewers", Proceedings of the 13th ACM SIGKDD, San Jose, California, pp. 500-509, 2007.

[4] Zhang, H., Giles, C.L., Foley, H.C., and Yen, J., "Probabilistic Community Discovery Using Hierarchical Latent Gaussian Mixture Model", Proceedings of 22nd AAAI Conference on Artificial Intelligence, Vancouver, British Columbia, Canada, pp. 663-668, July 22-26, 2007.

[5] Zhou, D., Manavoglu, E., Li, J., Giles, C.L., and Zha, H., "Probabilistic Models for Discovering Ecommunities", Proceedings of the World Wide Web (WWW), pp. 173-182, 2006.

[6] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z., "ArnetMiner: Extraction and Mining of Academic Social Networks", Proceedings of ACM SIGKDD, 2008.

[7] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P., "The Author-Topic Model for Authors and Documents", Proceedings of the 20th UAI, Banff, Canada, 2004.

[8] Blei, D.M., Ng, A.Y., and Jordan, M.I., "Latent Dirichlet Allocation", JMLR, Volume 3, pp. 993-1022, 2003.

[9] Andrieu, C., Freitas, N.D., Doucet, A., and Jordan, M., "An Introduction to MCMC for Machine Learning", Journal of Machine Learning, Volume 50, pp. 5-43, 2003.

[10] DBLP Bibliography Database. http://www.informatik.uni-trier.de/~ley/db/.

[11] Hofmann, T., "Probabilistic Latent Semantic Analysis", Proceedings of the 15th UAI, Stockholm, Sweden, 1999.

[12] Azzopardi, L., Girolami, M., and Risjbergen, K.V., "Investigating the Relationship between Language Model Perplexity and IR Precision-Recall Measures", Proceedings of the 26th ACM SIGIR, Toronto, Canada, 2003.

[13] Tyler, J.R., Wilkinson, D.M., and Huberman, B.A., "Email as Spectroscopy: Automated Discovery of Community Structure within Organizations", Proceedings of the C & T, pp. 81-96, 2003.

[14] Griffiths, T.L., and Steyvers, M., "Finding Scientific Topics", Proceedings of the NAS, pp. 5228-5235, USA, 2004.

[15] Newman, M.E., "Coauthorship Networks and Patterns of Scientific Collaboration", Proceedings of the National Academy, Volume 1, pp. 5200-5205, USA, 2004.

[16] Newman, M.E.J., "Fast Algorithm for Detecting Community Structure in Networks", Physical Review, Volume 69, pp. 066-133, 2004.

[17] Pothen, A., Simon, H., and Liou, K.P., "Partitioning Sparse Matrices with Eigenvectors of Graphs", SIAM Journal of SIMAX, Volume 11, pp. 430-452, 1990.

[18] Palla, G., Derenyi, I., Farkas, I., and Vicsek, T., "Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society", Nature, pp. 435-814, 2005.

[19] Girvan, M., and Newman, M.E.J., "Community Structure in Social and Biological Networks", Proceedings of the NAS, Volume 99, pp. 8271-8276, USA, 2002.

[20] Popescul, A., Flake, G.W., Lawrence, S., Ungar, L.H., and Giles, C.L., "Clustering and Identifying Temporal Trends in Document Databases", IEEE ADL, pp. 173-182, 2000.

[21] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D., "Dening and Identifying Communities in Networks", Proceedings of the NAS, USA, 2004.

[22] Ruan, J., and Zhang, W., "Identification and Evaluation of Weak Community Structures in Networks", Proceedings of the Association for the Advancement of Artificial Intelligence, 2006.

[23] Wilkinson, D.M., and Huberman, B.A.A., "Method for Finding Communities of Related Genes", Proceedings of the National Academy of Volume 1, pp. 5241-5248, USA, 2004.

[24] Breese, J., Heckerman, D., and Kadie, C., "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", Proceedings of the UAI, pp. 43-52, 1998.

[25] Deshpande, M,, and Karypis, G., "Item-Based Top-n-Recommendation Algorithms", ACM Transactions on Information Systems, Volume 22, No. 1, pp. 143-177, 2004.

[26] Balabanovic, M., and Shoham, Y., "Content-Based Collaborative Recommendation", Communications of the ACM, Volume 40, No. 3, 1997.

[27] Leskovec, J., Lang, K., Dasgupta, A., and Mahoney, M., "Statistical Properties of Community Structure in Large Social and Information Networks", Proceedings of the 17th WWW, 2008.

[28] Li, H., Nie, Z., Lee, W., Giles, C., and Rong, J., "Scalable Community Discovery on Textual Data with Relations", Proceedings of the 17th CIKM, 2008.

[29] Duan, D., Li, Y., Jin, Y., and Lu, Z., "Community Mining on Dynamic Weighted Directed Graphs", CNIKM Workshop, November 6, 2009.

[30] Lin, Y.R., Sun, J., Castro, P., Konuru, R., Sundaram, H., and Kelliher, A., "MetaFac: Community Discovery via Relational Hypergraph Factorization", Proceedings of ACM SIGKDD, June 28- July 1, 2009.

[31] Yang, T., Jin, R., Chi, Y., and Zhu, S., "Combining Link and Content for Community Detection: A Discriminative Approach", Proceedings of ACM SIGKDD, June 28- July 1, 2009.

[32] Zhou, D., Ji, X., Zha, H., and Giles, C.L., "Topic Evolution and Social Interactions: How Authors Effect Research", Proceedings of the CIKM, pp. 248-257, 2006.

[33] McCallum, A., Nigam, K., and Ungar, L.H., "Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching", Proceedings of the 6th ACM SIGKDD, pp. 169-178, 2000.