Conference Mining via Generalized Topic Modeling

Ali Daud[1], Juanzi Li[1], Lizhu Zhou[1], and Faqir Muhammad[2]

[1] Department of Computer Science & Technology, 1-308, FIT Building, Tsinghua University, Beijing, China, 100084
[2] Department of Mathematics & Statistics, Allama Iqbal Open University, Sector H-8, Islamabad, Pakistan, 44000
ali_msdb@hotmail.com, ljz@keg.cs.tsinghua.edu.cn, dcszlz@tsinghua.edu.cn, aioufsd@yahoo.com

Abstract. Conference Mining has been an important problem discussed these days for the purpose of academic recommendation. Previous approaches mined conferences by using network connectivity or by using semantics-based intrinsic structure of the words present between documents (modeling from document level (DL)), while ignored semantics-based intrinsic structure of the words present between conferences. In this paper, we address this problem by considering semantics-based intrinsic structure of the words present in conferences (*richer semantics*) by modeling from conference level (CL). We propose a generalized topic modeling approach based on Latent Dirichlet Allocation (LDA) named as Conference Mining (ConMin). By using it we can discover topically related conferences, conferences correlations and conferences temporal topic trends. Experimental results show that proposed approach significantly outperformed baseline approach in discovering topically related conferences and finding conferences correlations because of its ability to produce less sparse topics.

Keywords: Richer Semantics, Conference Mining, Generalized Topic Modeling, Unsupervised Learning.

1 Introduction

With the emergence of the Web, automatic acquirement of useful information from the text has been a challenging problem, when most of the information is implicit within the entities (e.g. documents, researchers, conferences, journals) and their relationships. For example, various conferences are held every year about different topics and huge volume of scientific literature is collected about conferences in digital libraries. It provides us with many challenging discovery tasks useful from researchers' point of view. For example, a new researcher can be interested in obtaining authoritative conferences of specific research area to do literature review or a group of researchers would like to know about conferences related to their research area for submitting papers.

Previous approaches used for conference mining problem can be categorized into two major frameworks 1) graph connectivity based approaches as a basis for representation and analysis of relationships between conferences [24,25] on the basis of co-authorship and publishing in the same venue and 2) topic modeling based approaches

W. Buntine et al. (Eds.): ECML PKDD 2009, Part I, LNAI 5781, pp. 244–259, 2009.

which make use of latent topic layer between words and documents to capture the semantic correlations between them. Recently one of the topic modeling approaches argued that conferences and authors are interdependent and should be modeled together [20]. Consequently, a unified topic modeling approach Author-Conference-Topic1 (ACT1) was proposed, which can discover topically related authors and conferences on the basis of semantics-based structure of the words by considering conferences information. Above mentioned frameworks based on graph connectivity ignored the semantics-based information. While, recent topic modeling approach viewed conferences information just as a stamp (token), which became the reason of ignoring implicit semantics-based text structure present between the conferences. We think this information is very useful and important for mining conferences.

In this paper, we will consider semantics-based text structure present between the conferences explicitly. We generalized previous topic modeling approach [20] idea of mining conferences from a single document "Constituent-Document" (*poorer semantics* because of only some semantically related words are present in one document) to all publications of conference "Super-Document" (*richer semantics* because of many semantically related words are present in all documents of one conference). It can provide grouping of conferences in different groups on the basis of latent topics (semantically related probabilistic cluster of words) present between the conferences. We propose a Latent Dirichlet Allocation (LDA) [4] based ConMin approach which can discover topically related conferences. We used discovered topics to find associations between conferences by using sKL divergence and shown temporal topic trends of conferences. We empirically showed that ConMin approach clearly achieve better results than ACT1 approach for conference mining and solution provided by us produced quite intuitive and functional results.

The novelty of work described in this paper lies in the; formalization of the key conference mining issues, proposal of generalized topic modeling (ConMin) approach to deal with the issues by capturing *richer semantics*, and experimental verification of the effectiveness of our approach on real-world dataset. To the best of our knowledge, we are the first to deal with the aforementioned conference related discovery issues directly (not through authors generated topics like ACT1) by proposing a generalized topic modeling approach from DL to CL.

The rest of the paper is organized as follows. In Section 2, we formalize the key conference related mining issues. Section 3 illustrates our proposed approach for modeling conferences with its parameter estimation details. In Section 4, dataset, parameters settings, performance measures, baseline approach with empirical studies and discussions about the results are given; applications of proposed approach are provided at the end of this section. Section 5 provides related work and section 6 brings this paper to the conclusions and future work.

Note that in the rest of the paper, we use the term constituent-document, accepted paper, and document interchangeably. Additionally "super-document" means all the documents of one conference.

2 Problem Setting

Our work is focused on mining conferences through their accepted papers. Each conference accepts many papers every year. To our interest, each publication contains

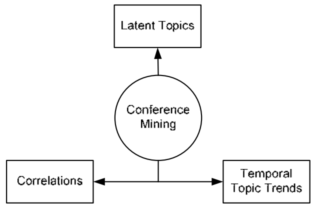

Fig. 1. Conferences related discovery issues

title which covers most of the highly related sub research areas. Conferences with their accepted papers on the basis of latent topics can be mined. Figure 1 provides a pictorial look of conference related mining issues discussed here.

We denote a conference (Super-Document) c as a vector of N_c words based on all accepted papers (Constituent-Documents) by the conference and formalize conference mining problem as three subtasks. Intuition behind considering conference as super-document is based on thinking that semantics at super-document level are richer as compared to semantics at a single document (Constituent-Document).

1) Discovery and Ranking of Conferences related to Topics: Given a conference c with N_c words, find the latent topics Z of conference. Formally for a conference, we need to calculate the probability $p(z|c)$, where z is a latent topic and c is a conference.
   Predict Z topics for a conference: Given a new conference c (not contained previously in the corpus) with W_c words, predict the topics contained in the conference.
2) Discovery of Conferences Correlations: Given two conferences c_1 and c_2 with N_{c1} and N_{c2} words respectively, find the correlations between conferences.
3) Discovery of Conferences Temporal Topic Trends: Given a conference c with N_c words for every year, access the temporal topic likeliness of a conference.

3 Conference Modeling

In this section, before describing our ConMin approach, we will first describe how documents are modeled with topics using topic model LDA, followed by modeling of conferences with authors' topics (ACT1 approach).

3.1 Modeling Documents with Topics (LDA)

Fundamental topic modeling assumes that there is a hidden topic layer $Z = \{z_1, z_2, z_3, \ldots, z_i\}$ between the word tokens and the 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. LDA [4] is a state-of-the-art topic modeling approach which makes use of latent topic layer to capture semantic dependencies between the words. 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:

$$P(w|d,\theta,\emptyset) = \sum_{z=1}^{T} P(w|z,\emptyset_z)P(z|d,\theta_d) \tag{1}$$

3.2 Modeling Conferences with Authors Topics (ACT1 (DL) Approach)

Recently, LDA is extended to discover topically related conferences indirectly by using topics of documents generated by authors [20]. In ACT1 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 generative probability of the word *w* with conference *c* for author *r* of a document *d* is given as:

$$P(w,c|r,d,\emptyset,\Psi,\theta) = \sum_{z=1}^{T} P(w|z,\emptyset_z)P(c|z,\Psi_z)P(z|r,\theta_r) \tag{2}$$

3.3 Modeling Conferences with Topics (ConMin (CL) Approach)

The basic idea of topic modeling that words and documents can be modeled by considering latent topics became the intuition of modeling the words and conferences directly through latent topics. We generalize this idea from DL [4] to CL by considering documents as sub-entities of a conference. In our approach a conference is viewed as a composition of the words of its all accepted publications. Symbolically, for a conference *c* we can write it as: $C = \{\mathbf{d_1} + \mathbf{d_2} + \mathbf{d_3} + \ldots + \mathbf{d_i}\}$, where d_i is one document in a conference.

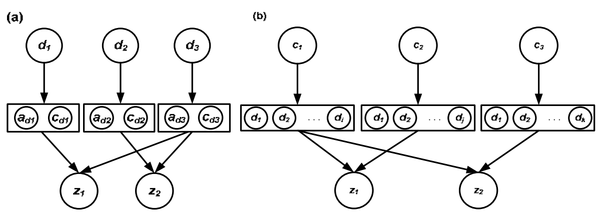

Fig. 2. Conference modeling a) ACT1 (DL) and b) ConMin (CL) approaches

DL approach is responsible for generating latent topics of documents, while CL approach is responsible for generating latent topics of conferences. For each conference *c*, a multinomial distribution θ_c over topics is randomly sampled from a Dirichlet with parameter α, and then for each word *w* contained in super-document, 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 with parameter β.

The generative process is as follows:

1. For each conference $c = 1,\ldots, C$
   Choose θ_c from Dirichlet (α)

2. For each topic $z = 1,\ldots, T$
   Choose Φ_z from Dirichlet (β)
3. For each word $w = 1,\ldots, N_c$ of conference c
   Choose a topic z from multinomial (θ_c)
   Choose a word w from multinomial (Φ_z)

Figure 3 shows the generating probability of the word w from the conference c is given as:

$$P(w|c,\theta,\emptyset) = \sum_{z=1}^{T} P(w|z,\emptyset_z)P(z|c,\theta_c) \tag{3}$$

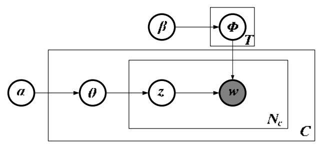

Fig. 3. ConMin approach (generalized smoothed LDA)

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

$$P(z_i = j|\mathbf{z}_{-i},\mathbf{w}) \propto \frac{n_{-i,j}^{(wi)}+\beta}{n_{-i,j}^{(\cdot)}+W\beta}\frac{n_{-i,j}^{(ci)}+\alpha}{n_{-i,\cdot}^{(ci)}+Z\alpha} \tag{4}$$

where $z_i = j$ represents the assignments of the i^{th} word in a conference to a topic j. $\mathbf{z}_{-i}$ represents all topic assignments excluding the i^{th} word, and $\mathbf{w}$ represents all words in the dataset. Furthermore, $n_{-i,j}^{(wi)}$ is the total number of words associated with topic j, excluding the current instance, and $n_{-i,j}^{(ci)}$ is the total number of words from conference c assigned to topic j, excluding the current instance. "." 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.

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

$$\emptyset_{zw} = \frac{n_{-i,j}^{(wi)}+\beta}{n_{-i,j}^{(\cdot)}+W\beta} \tag{5}$$

$$\theta_{cz} = \frac{n_{-i,j}^{(ci)}+\alpha}{n_{-i,\cdot}^{(ci)}+Z\alpha} \tag{6}$$

where, $\emptyset_{zw}$ is the probability of word w in topic z and θ_{cz} is the probability of topic z for conference c. These values correspond to the predictive distributions over new words w and new topics z conditioned on w and z.

4 Experiments

4.1 Dataset

We downloaded five years publication dataset of conferences from DBLP [8,14] by only considering conferences for which data was available for years 2003-2007. In total, we extracted 90,124 publications for 261 conferences and combined them into a super-document separately for each conference. We then preprocessed corpus by a) removing stop-words, punctuations and numbers b) down-casing the obtained words, and c) removing words that appear less than three times in the corpus. This led to a vocabulary size of V=10,902 and a total of 571,439 words in the corpus. Figure 4 shows quite smooth yearly data distribution for number of publications in the conferences.

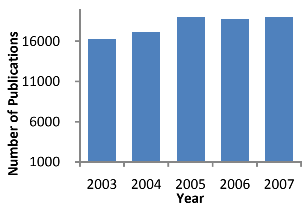

Fig. 4. Histogram illustrating data distribution

4.2 Parameter Settings

One can estimate the optimal values of hyper-parameters α and β (figure 3) by using Expectation Maximization (EM) method [11] or Gibbs sampling algorithm [10]. EM algorithm is susceptible to local maxima and computationally inefficient [4], consequently Gibbs sampling algorithm is used. For some applications topic models are sensitive to the hyper parameters and need to be optimized. For application in this paper, we found that our topic model based approach is not sensitive to the hyper parameters. In our experiments, for 200 topics Z the hyper-parameters α and β were set at 50/Z and .01 respectively. The numbers of topics Z were fixed at 200 on the basis of human judgment of meaningful topics plus measured perplexity [2] on 20% held out test dataset for different number of topics Z from 2 to 300. We ran five independent Gibbs sampling chains for 1000 iterations each. All experiments were carried out on a machine running Windows XP 2006 with AMD Athlon I Dual Core Processor (1.90 GHz) and 1 GB memory. The run time per each chain was 1.26 hours.

4.3 Performance Measures

Perplexity is usually used to measure the performance of latent-topic based approaches; however it cannot be a statistically significant measure when they are used for information retrieval [Please see [2] for details]. In our experiments, at first we used average entropy to measure the quality of discovered topics, which reveals the purity of topics. Entropy is a measure of the disorder of system, less intra-topic entropy is usually better. Secondly, we used average Symmetric KL (sKL) divergence [19] to measure the quality of topics, in terms of inter-topic distance. sKL divergence is used here to measure the relationship between two topics, more inter-topic sKL divergence (distance) is usually better.

To measure the performance in terms of precision and recall [2] 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 conferences ranking. We discovered top 9 conferences related to top most conference (e.g. for ConMin "XML Databases" topic it is XSym) in each topic by using sKL divergence [please see table 1]. We compared these top 9 conferences with topically discovered top 10 conferences and calculated error rate with respect to their absence or presence in the topically ranked conferences list.

$$Entropy\ of\ (Topic) = -\sum_z P(z) log_2[P(z)] \tag{7}$$

$$sKL(i,j) = \sum_{z=1}^{T}\left[\theta_{iz} log\frac{\theta_{iz}}{\theta_{jz}} + \theta_{jz} log\frac{\theta_{jz}}{\theta_{iz}}\right] \tag{8}$$

4.4 Baseline Approach

We compared proposed ConMin with ACT1 and used same number of topics for comparability. The numbers of Gibbs sampler iterations used for ACT1 are 1000 and parameter values same as the values used in [20]. We used the same machine which was used for proposed approach; run time per each chain for ACT1 was 3.00 hours almost double than proposed approach. It shows that ConMin approach is also better in terms of time complexity.

4.5 Results and Discussions

The effect of topic sparseness on the model performance is studied both qualitatively and quantitatively. Firstly, we provide qualitative comparison between ConMin and ACT1 approaches. We discovered and probabilistically ranked conferences related to specific area of research on the basis of latent topics. Table 1 illustrates 7 different topics out of 200, discovered from the 1000th iteration of a particular Gibbs sampler run. The words associated with each topic for ConMin approach are strongly semantically related (less sparse) than that of ACT1, as they are assigned higher probabilities (please see prob. column in table 1). So, they make compact topics in the sense of conveying a semantic summary of a specific area of research [Please see figure 5 to see quantitative comparison of topic compactness]. Additionally it is observed that because of topic sparseness topically related conferences are also sparse (not from the specific area of research).

Consequently the conferences associated with each topic for ConMin are also more precise than ACT1, as they are assigned high probabilities (please see prob. Column in table 1). Only higher probabilities assigned to topic words and conferences is not extremely convincing, so we also investigated the bad impact of topic sparseness due to lower probabilities on the performance of baseline approach. For example, from top ten conferences six conferences related to "XML Databases" topic discovered by ACT1 are VLDB, SIGMOD, ICDE, Xsym, ADBIS, WIDM which are related to databases research area and other four ECOOP, SEKE, CAISE and KI are more related to software engineering and artificial intelligence research areas. While for ConMin topic "XML Databases" all the conferences are related to only databases research area. Similarly for "Data Mining" topic top ten conferences discovered by ConMin are more precise then ACT1 as for ACT1 SAC (Cryptography), CCGRID (Cluster Computing and Grid), ACM SenSys (Embedded Networked and Sensor Systems), ICDCS (Distributed Computing Systems) and ISISC (Information Security and Cryptology) are not actually related to data mining research area, additionally ACT1 is unable to find PAKDD, PKDD, DAWAK and DS for "Data Mining" topic among top ten conferences but they are well-known conferences in this field. One can see that PKDD and PAKDD are discovered by ACT1 for "Web Search" topic, which mismatches with the real world data. Similar kind of problem is encountered by ACT1 for other topically related conferences. It concludes that sparser the topics the discovered conferences will also be sparse which will result in poor performance of the approach.

Here it is obligatory to mention that top 10 conferences associated with a topic are not necessarily most well-known conferences in that area, but rather are the conferences that tend to produce most words for that topic in the corpus. However, we see that top ranked conferences for different topics are in fact top class conferences of that area of research for proposed approach. For example for topic 28 "Bayesian Networks" and topic 117 "XML Databases" top ranked conferences are more or less the

Table 1. An illustration of 7 discovered topics (top ConMin approach, bottom ACT1 approach). Each topic is shown with the top 10 words and conferences. The titles are our interpretation of the topics.

Topic 117 (ConMin)		**Topic 164 (ConMin)**		**Topic 63 (ConMin)**		**Topic 138 (ConMin)**		**Topic 190 (ConMin)**		**Topic 28 (ConMin)**		**Topic 0 (ConMin)**	
"XML Databases"		**"Semantic Web"**		**"Information Retrieval"**		**"Digital Libraries"**		**"Data Mining"**		**"Bayesian Networks"**		**"Web Search"**	
Word	***Prob.***	***Word***	***Prob.***	***Word***	***Probability***	***Word***	***Prob.***	***Word***	***Prob.***	***Word***	***Prob.***	***Word***	***Prob.***
xml	0.121514	semantic	0.125522	retrieval	0.157699	digital	0.234255	mining	0.147924	Bayesian	0.083057	web	0.328419
query	0.059027	web	0.12249	information	0.112182	libraries	0.099236	data	0.107059	networks	0.057923	search	0.02874
databases	0.0547	owl	0.03093	query	0.05448	library	0.09544	clustering	0.056024	inference	0.042624	content	0.024066
database	0.052969	rdf	0.029718	relevance	0.037277	metadata	0.031998	frequent	0.044513	time	0.028964	semantic	0.024066
processing	0.050199	ontologies	0.023048	feedback	0.029392	access	0.020611	patterns	0.036455	belief	0.028418	xml	0.019565
queries	0.045179	annotation	0.01941	search	0.022583	collections	0.01573	time	0.027054	causal	0.024593	language	0.018007
relational	0.03237	end	0.016378	user	0.020074	collection	0.013019	streams	0.02667	continuous	0.0235	pages	0.017314
efficient	0.025447	data	0.01274	language	0.017924	image	0.012477	pattern	0.022066	graphical	0.022954	information	0.015929
management	0.020773	large	0.010921	xml	0.017565	educational	0.012477	high	0.021298	structured	0.021315	user	0.014717
schema	0.020081	networks	0.010921	term	0.017207	oai	0.011935	privacy	0.017077	graphs	0.019676	collaborative	0.014544
Conference	***Prob.***	***Conference***	***Prob.***	***Conference***	***Prob.***	***Conference***	***Prob.***	***Conference***	***Prob.***	***Conference***	***Prob.***	***Conference***	***Prob.***
Xsym	0.413636	ISWC	0.330486	SIGIR	0.242417	JCDL	0.293113	SDM	0.251071	UAI	0.227882	WWW	0.234292
VLDB	0.199081	ASWC	0.326289	ECIR	0.194643	ECDL	0.27024	KDD	0.213337	AAAI	0.049531	LA-WEB	0.214421
SIGMOD	0.197517	WWW	0.040461	CIKM	0.086882	ELPBU	0.086239	ICDM	0.198849	NIPS	0.048314	WISE	0.213057
ICDE	0.192734	WIDM	0.014888	SPIRE	0.053974	MKM	0.04002	PKDD	0.196895	ICML	0.046224	WIDM	0.192592
IDEAS	0.1875	PODS	0.01374	SEBD	0.037998	DOCENG	0.025634	PAKDD	0.187208	ECML	0.044391	ICWS	0.159733
ADBIS	0.179348	ICCS	0.010382	ECDL	0.036844	Hypertext	0.017996	DAWAK	0.15004	Cana. AI	0.030308	WI	0.157155
SEBD	0.17217	ACSAC	0.009259	MMM	0.032828	SBBD	0.012186	DS	0.072158	ICTAI	0.016417	Hypertext	0.114341
BNCOD	0.165171	CAISE	0.008955	ICWS	0.029954	ECOOP	0.010417	IDEAS	0.066027	SDM	0.016065	ICWL	0.09839
ADC	0.164414	PSB	0.00837	WAIM	0.027234	SIGCSE	0.008574	ICDE	0.0647	EC	0.014017	ICWE	0.073778
PODS	0.162534	CADE	0.008267	ELPBU	0.022441	ECIR	0.008135	SSDBM	0.061772	AUSAI	0.012357	ASWC	0.0631
Topic 117 (ACT1)		**Topic 164 (ACT1)**		**Topic 63 (ACT1)**		**Topic 138 (ACT1)**		**Topic 190 (ACT1)**		**Topic 28 (ACT1)**		**Topic 0 (ACT1)**	
"XML Databases"		**"Semantic Web"**		**"Information Retrieval"**		**"Digital Libraries"**		**"Data Mining"**		**"Bayesian Networks"**		**"Web Search"**	
Word	***Prob.***	***Word***	***Prob.***	***Word***	***Probability***	***Word***	***Prob.***	***Word***	***Prob.***	***Word***	***Prob.***	***Word***	***Prob.***
data	0.03135	semantic	0.056959	retrieval	0.035258	digital	0.056555	data	0.029013	Bayesian	0.017148	web	0.065414
xml	0.031176	web	0.05335	information	0.020689	libraries	0.026451	mining	0.021635	learning	0.01287	search	0.017745
query	0.023387	ontology	0.025683	search	0.018469	library	0.021862	clustering	0.020054	networks	0.011704	based	0.016747
database	0.01802	based	0.016861	based	0.016387	based	0.012868	patterns	0.008459	models	0.010926	semantic	0.015748
web	0.013	ontologies	0.012851	web	0.015277	information	0.00938	learning	0.007668	inference	0.006649	services	0.007512
system	0.012135	owl	0.011247	text	0.01167	metadata	0.008279	based	0.007668	probabilistic	0.00626	data	0.006514
processing	0.011789	services	0.010846	document	0.011392	evaluation	0.006994	classification	0.007141	based	0.005871	information	0.006514
based	0.011096	rdf	0.010045	query	0.010976	web	0.00681	preserving	0.006351	markov	0.004705	approach	0.005765
relational	0.010231	approach	0.008842	relevance	0.009588	collections	0.00681	streams	0.006087	graphical	0.004705	queries	0.005765
management	0.010231	service	0.008441	evaluation	0.007646	search	0.006627	privacy	0.005824	information	0.004705	query	0.005516
Conference	***Prob.***	***Conference***	***Prob.***	***Conference***	***Prob.***	***Conference***	***Prob.***	***Conference***	***Prob.***	***Conference***	***Prob.***	***Conference***	***Prob.***
VLDB	0.450054	ASWC	0.496074	SIGIR	0.651289	JCDL	0.609793	SDM	0.695489	UAI	0.978935	WWW	0.986798
SIGMOD	0.378506	ISWC	0.49582	ECIR	0.249118	ECDL	0.379536	ICDM	0.185296	NIPS	0.001382	CIKM	0.001388
ICDE	0.150949	ICWS	0.000534	CIKM	0.080613	WISE	0.00207	KDD	0.102877	ISAAC	0.000724	ECIR	0.000711
Xsym	0.014415	KI	0.00028	SPIRE	0.014316	SBBD	0.00116	VLDB	0.002225	AUSAI	0.000724	PKDD	0.000711
ECOOP	0.000233	IEAAIE	0.00028	DAWAK	0.000179	SODA	0.000705	ICDE	0.00186	PODS	0.000724	SPIRE	0.000711
SEKE	0.000233	INFOCOM	0.00028	PKDD	0.000179	DOCENG	0.00025	SAC	0.000766	SIGIR	0.000724	TCVG	0.000372
WIDM	0.000233	LA-WEB	0.00028	WISE	0.000179	CASES	0.00025	CCGRID	0.000766	AINA	0.000066	TAB. AUX	0.000372
CAISE	0.000021	ADBIS	0.000025	KI	0.000016	ACL	0.00025	SenSys	0.000401	CAISE	0.000066	PAKDD	0.000372
KI	0.000021	AGILE	0.000025	ADBIS	0.000016	ECOOP	0.00025	ICDCS	0.000401	KI	0.000066	KI	0.000034
ADBIS	0.000021	XP	0.000025	AGILE	0.000016	CAISE	0.000023	ISISC	0.000401	ADBIS	0.000066	ADBIS	0.000034

best conferences of artificial intelligence and databases fields, respectively. Both topics also show deep influence of Bayesian networks on artificial intelligence and move from simple databases to XML database, respectively. We think, characteristically in top class conferences submitted papers are very carefully judged for the relevance to the conference research areas which results in producing more semantically related words; this is why top class conferences are ranked higher.

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 dataset. A fraction of non-research topics, perhaps 10-15%, are also discovered that are not directly related to a specific area of research, as the words present in those topics were actually used as a glue between scientific terms. In addition to qualitative comparison between ConMin and ACT1, we also provide quantitative comparison to explain the effect of topics sparseness on the performance of approach. Figure 5 (a) shows the average entropy of topic-word distribution for all topics measured by using equation 7. Lower entropy curve of proposed approach for different number of topics Z = 50, 100, 150, 200, 250, 300 shows its effectiveness for obtaining less sparse topics which resulted in its better ranking performance shown in table 1. Figure 5 (b) shows the average distance of topic-word distribution between all pairs of the topics measured by using equation 8. Higher sKL divergence curve for different number of topics Z = 50, 100, 150, 200, 250, 300 confirms the effectiveness of the proposed approach for obtaining compact topics as compared to baseline approach.

From the curves in figure 5 (a) and figure 5 (b) it is clear that ConMin approach outperformed ACT1 approach for different number of topics. The performance difference for different number of topics is pretty much even, which corroborate that proposed approach dominance is not sensitive to the number of topics.

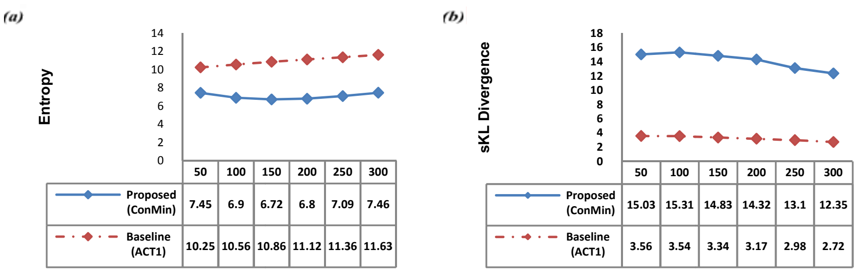

Fig. 5. a) Average Entropy curve as a function of different number of topics, lower is better and b) Average sKL divergence curve as a function of different number of topics, higher is better

Now we provide comparison in terms of error rate. Table 2 shows top 9 conferences discovered related to the first conference of each topic for ConMin and ACT1 approaches by using sKL divergence. For example, in case of "XML Databases" topic ADC, ADBIS, IDEAS, BNCOD, VLDB, SIGMOD, PODS, DASFAA and DEXA are top 9 conferences correlated with "Xsym" for ConMin approach.

The highlighted blocks in table 2 shows that similar results are found for discovered topics in table 1 and sKL divergence calculated for top most conference. For example, in case of ConMin approach top 10 conferences shown in table 1 for "XML Databases" topic has 7 conferences in common, which are ADC, ADBIS, IDEAS, BNCOD, VLDB, SIGMOD and PODS. From top 9 related conferences for seven selected topics (same is the case with non selected topics) shown in the table 2 the error rate (ER) for ConMin is less than ACT1, except digital libraries topic and ConMin approach has 30.16 % less average error rate than ACT1. It shows the bad effect of topics sparseness on conferences ranking performance of ACT1, and its inability to discover better results in comparison with proposed approach.

Table 2. An illustration of 7 topics sparseness effect on ranking in terms of error rate (ER). Here acronyms are XML Databases (XMLDB), Semantic Web (SeW), Information Retrieval (IR), Digital Libraries (DiL), Data Mining (DM), Bayesian Networks (BN) and Web Search (WS).

ConMin Approach							**ACT1 Approach**						
XMLDB	**SeW**	**IR**	**DiL**	**DM**	**BN**	**WS**	**XMLDB**	**SeW**	**IR**	**DiL**	**DM**	**BN**	**WS**
ADC	ASWC	ECIR	ECDL	ICDM	ICML	WI	SIGMOD	ISWC	ECIR	ECDL	KDD	EC	Hypertext
ADBIS	ER	CIKM	ELPBU	PAKDD	ECML	LA-WEB	ICDE	LA-WEB	CIKM	WISE	ICDM	ICML	SPIRE
IDEAS	LA-WEB	NLDB	Hypertext	KDD	NIPS	WISE	Xsym	KI	SPIRE	SBBD	SEDB	ALT	LISA
BNCOD	ISTA	ACL	WWW	PAKDD	AAAI	ICWS	ADA	ADA	WISE	ISI	ICDE	PODS	MATES
VLDB	WI	ICWS	ICWL	DS	ALT	CIKM	Ada-Eu	Xsym	MKM	ECOOP	VLDB	ADA	SGP
SIGMOD	SEBD	WWW	SIGIR	ECML	COLT	WAIM	ISTA	PPDP	DOCENG	DOCENG	ISISC	COLT	ICSOC
PODS	WWW	WISE	DOCENG	DAWAK	Cana. AI	WIDM	SDM	FSTCS	TableAUX	SODA	ADA	ISAAC	SIGIR
DASFAA	CAISE	KDD	ECIR	IDEAL	SDM	Hypertext	ICFP	ECOOP	ISSAC	CASES	SAC	Xsym	ICWS
DEXA	WIDM	MMM	LA-WEB	ICML	ICTAI	JCDL	APLAS	ICWS	RCLP/LPAR	ADA	SAM	PPDP	FC
ER=22.22	ER=55.55	ER=55.55	ER=44.44	ER=33.33	ER=22.22	ER=33.33	ER=66.66	ER=66.66	ER=55.55	ER=33.33	ER=33.33	ER=77.77	ER=88.88
Average Error Rate = 30.15							**Average Error Rate = 60.31**						

4.6 Applications of Proposed Approach

4.6.1 Topics for New Conferences

One would like to quickly access the topics for new conferences which are not contained in the training dataset by offline trained model. Provided parameter estimation Gibbs sampling algorithm requires significant processing time for large number of conferences. It is computationally inefficient to rerun the Gibbs sampling algorithm for every new conference added to the dataset. For this purpose we apply equation 4 only on the word tokens in the new conference each time temporarily updating the count matrices of (word by topic) and (topic by conference). The resulting assignments of words to topics can be saved after a few iterations (20 in our simulations which took only 2 seconds for one new conference). Table 3 shows this type of inference. To show predictive power of our approach we treated two conferences as test conferences one at a time, by training model on remaining 260 conferences to discover latent topics. Discovered topics are then used to predict the topics for words of the test conference.

Predicted words associated with each topic are quite intuitive, as they provide a summary of a specific area of research and are true representatives of conferences. For example, KDD conference is one of the best conferences in the area of Data Mining. Top five predicted topics for this conference are very intuitive, as "Data Mining", "Classification and Clustering", "Adaptive Event Detection", "Data Streams" and "Time Series Analysis" all are prominent sub-research areas in the field of data mining and knowledge discovery. Topics predicted for SIGIR conference are also intuitive and precise, as they match well with conference sub-research areas. Comparatively ACT1 (DL) approach is unable to directly predict topics for new conferences.

Table 3. An illustration of top five predicted topics for SIGIR and KDD conferences; each topic is shown with its probability, title (our interpretation of the topics) and top 10 words

SIGIR		
Topic Words	**Title**	**Probability**
retrieval, search, similarity, query, based, clustering, classification, relevance, document, evaluation	Information Retrieval	.2001
information, based, text, document, approach, documents, web, user, content, structured	Web based Information	.1340
language, text, extraction, semantic, disambiguation, question, word, answering, relations, natural	Intelligent Question Answering	.0671
web, search, collaborative, xml, user, pages, information, mining, content, sites	Web Search	.0415
models, probabilistic, random, structure, graph, exploiting, conditional, hidden, probability, markov	Probabilistic Models	.0361
KDD		
Topic Words	**Title**	**Probability**
mining, clustering, data, patterns, discovery, frequent, association, rules, algorithm, rule	Data Mining	.1819
classification, data, feature, selection, clustering, support, vector, machine, machines, Bayesian	Classification and Clustering	.0809
based, approach, model, multi, algorithm, method, efficient, analysis, detection, adaptive	Adaptive Event Detection	.0652
data, streams, stream, similarity, semantic, queries, incremental, adaptive, distributed, trees	Data Streams	.0618
time, high, large, efficient, dimensional, series, method, scalable, correlation, clusters	Time Series Analysis	.0584

In addition to the quantitative and qualitative evaluation of topically related conferences, we also quantitatively illustrate the predictive power of proposed approach in predicting words for the new conferences. For this purpose, perplexity is derived for conferences by averaging results for each conference over five Gibbs samplers. The perplexity for a test set of words W_c, for conference c of test data C_{test} is defined as:

$$perplexity\ (C_{test}) = exp\left[-\frac{logp(W_c)}{N_c}\right] \tag{9}$$

Figure 6 shows the average perplexity for different number of topics for AAAI, SIGIR, KDD and VLDB conferences, which fairly indicate the stable predictive power of proposed approach after 50 topics for all conferences.

4.6.2 Conference Correlations

ConMin and ACT1 both approaches can be used for automatic correlation discovery [19] between conferences, which can be utilized to conduct joint conferences in the future. To illustrate how it can be used in this respect, distance between conferences i and j is calculated by using equation 8 for topics distribution conditioned on each of the conferences distribution.

We calculated the dissimilarity between the conferences by using equation 8, smaller dissimilarity values means higher correlation between the conferences. For similar pairs less dissimilarity value and for dissimilar pairs higher dissimilarity value indicate better performance of our approach.

Table 4 shows correlation between 8 pairs of conferences, with every two pairs in order from top to down have at least one conference in common making four (A, B, C, D) common pairs. Common conference pairs show the effectiveness of our approach in discovering more precise conferences correlations. For example, common pair A has ASWC (Asian Semantic Web Conference) conference common in pairs (1, 2). Dissimilarity value between pair 1 (pretty much related conferences Asian Semantic Web Conference and International Semantic Web Conference) is smaller for ConMin .176 than that of ACT1 2.75, and dissimilarity value between pair 2 (related conferences to normal extent) is smaller for ConMin 3.16 than that of ACT1 3.61, which shows that ConMin can find correlations better. Common pair B has ECIR (European Conference on Information Retrieval) common in pairs (3, 4). Dissimilarity value between pair 3 is smaller for ConMin 1.13 than that of ACT1 1.89 because both are IR related conferences, while dissimilarity value between pair 4 is greater for ConMin 4.03 than that of ACT1 1.58 because ECIR is top ranked conference for IR topic in table 1 and JCDL (Joint Conference on Digital Libraries) is top ranked conference for topic Digital Libraries in table 1 for both approaches, which shows that ConMin can better disambiguate which conference is related to which conference and to which extent. On the other hand according to ACT1 approach ECIR is more related to JCDL 1.58 than SIGIR (Special Interest Group Conference on Information Retrieval) 1.89 which is against the real world situation. The results for pairs C and D represent same situation as pair B, which proves overall authority of ConMin approach on ACT1 in capturing semantics-based correlations between conferences.

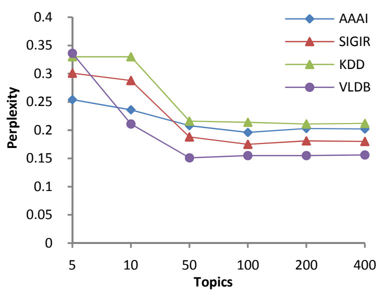

Fig. 6. Measured perplexity for new conferences

Table 4. *sKL* divergence for pairs of Conferences of ConMin and ACT1

Common Pairs	Pairs	Conferences	T=200 ConMin	T=200 ACT1
A	1	ASWC	.176	2.75
		ISWC		
	2	ASWC	3.16	3.61
		WWW		
B	3	ECIR	1.13	1.89
		SIGIR		
	4	ECIR	4.03	1.58
		JCDL		
C	5	SDM	1.49	2.31
		KDD		
	6	SDM	3.91	1.25
		UAI		
D	7	PODs	2.28	3.33
		VLDB		
	8	PODs	7.68	3.16
		ISWC		

4.6.3 Conferences Temporal Topic Trends

In most of the cases, conferences can be dominated by different topics in different years, which can provide us with topic drift for different research areas in different conferences. We used yearly data from (2003-2007) to analyze these temporal topic trends. Using 200 topics *Z*; for each conference corpus was partitioned by year, and for each year all of the words were assigned to their most likely topic using ConMin approach. It provided us the probability of topics assigned to each conference for a given year. The results provide interesting and useful indicators of temporal topic status of conferences. Figure 7 shows the results of plotting topics for SIGIR and KDD, where each topic is indicated in the legend with the five most probable words. Temporal conference trends can be captured by Topics over Time [22] and Dynamic Topic Models [5], but we are not focusing on that here.

The left plot shows the super dominant continuing topic "Information Retrieval" and other four topics having very low and steady likeliness trend for SIGIR conference. The right plot shows the ongoing dominancy of "Data Mining" topic and steady increase in the popularity of topics "Information Retrieval" and "Vector based Learning" for KDD (Knowledge Discovery in Databases) conference. As a whole, both conferences are dominated by one topic over the years, which is also one of the

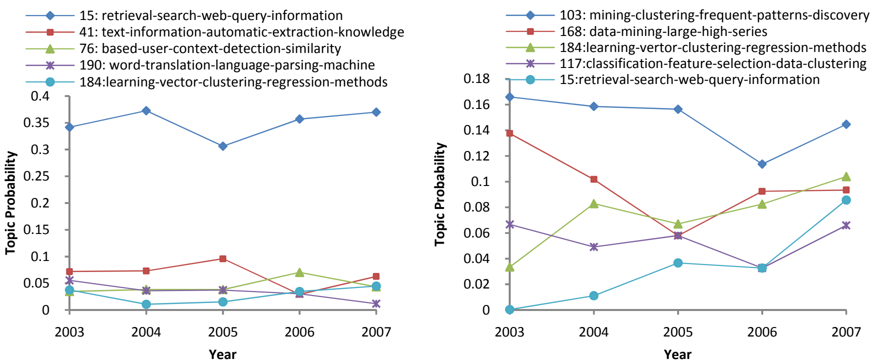

Fig. 7. Temporal topic trends of conferences

judgment criteria of the excellence of the conference and ongoing popularity of that topic. Here, it is necessary to mention that the probability for each topic per year of a conference only indicates probabilities assigned to topics by our approach, and makes no direct assessment of the quality or importance of the particular sub-area of a conference. Nonetheless, despite these caveats, obtained results are quite informative and indicate understandable temporal status of research topics in the conferences. Comparatively, ACT1 (DL) approach is unable to directly discover temporal topic trends.

5 Related Work

Automatic extraction of topics from text is performed by [15,16] to cluster documents into groups based on similar semantic contents. Clustering provides a good way to group similar documents, but clustering is inherently limited by the fact that each document is only associated with one cluster. For this reason soft clustering representation techniques are mandatory, which can allow documents composed of multiple topics to relate to more than one cluster on the basis of latent topics.

Probabilistic Latent Semantic Indexing (PLSI) [11] was proposed as a probabilistic alternative to projection and clustering methods. While PLSI produced impressive results on a number of document modeling problems, the number of parameters in the model grows linearly with the size of the corpus, which leads to serious problems of model over fitting and it was not clear how to assign a probability to a document outside the corpus.

Consequently, a more general probabilistic topic model LDA was proposed [4]. LDA assumes that each word in the document is generated by a latent topic and explicitly models the words distribution of each topic as well as the prior distribution over topics in the document. We generalized LDA to model conferences directly instead of indirectly modeling conferences like ACT1 [20] which modeled conferences through topics generated by the authors, and obtained more precise results.

Entities are modeled as graphs and related groups of entities were discovered either by network linkage information [17] or by iterative removal of edges between graphs [9,18,21]. Collaborative filtering [6,7] 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 used to recommend items on the basis of correlations between the content of the items and the users' preferences. This method creates a profile for each item or user to characterize their nature.

Previously, topics of conferences are extracted on the basis of keyword frequency from paper titles for related conferences finding [24] and specific area conferences are suggested by using pair-wise random walk algorithm [25], without considering semantic information present in the text. Differently, a topic modeling approach is used to discover topically related conferences [20]. Aforementioned approaches were incapable of considering implicit semantic information based text structure present between conferences. While, in real world co-occurrence of words and conferences; instead of co-occurrence of words and documents, can provide more appropriate semantics-based conferences correlations.

Traditionally, Kernighan-Lin algorithm and spectral bisection method [12,17] used the network linkage information between the entities to find the relationships between them. Both approaches are useless if there is no network connectivity information. Differently, correlations between authors and topics are discovered by using semantic information presented in the text [19]. Recently, Eclipse Developers correlations are discovered by using KL Divergence [13]. Here, we used sKL Divergence to discover semantics-based correlations between conferences.

Temporal topic trends of computer science were discovered in Citeseer documents [16,19] by utilizing clustering and semantics-based text information. Recently, Dynamic Topic model and Topics over Time [5,22] are used to find the general topic trends in the field of computer science. A Bayesian Network was proposed on the basis of authors to understand the research field evolution and trends [23]. Here, we used ConMin to discover topic trends specific to conferences without using authors' information, these topics are also representative of general topic trends in computer science field.

6 Conclusions and Future Work

This study deals with the problem of conference mining through capturing rich semantics-based structure of words present between conferences. We conclude that our generalization from DL to CL is significant; as proposed generalized approach's discovered and probabilistically ranked conferences (can also be applied to journals datasets such as HEP or OHSUMED) related to specific knowledge domains are better than baseline approach. While, predicted topics for new conferences are practical and meaningful. Proposed approach was also proved effective in finding conferences correlations when compared with the baseline approach. We demonstrated the effectiveness of proposed approach by applying it for analyzing temporal topic trends, which provide useful information. CL (capturing conference-level semantic structure) approach can handle the problem of DL (not capturing conference-level semantic structure) approach and provides us with dense topics. We studied the effect of generalization on topics denseness and concluded that sparser topics will results in poor performance of the approach. Empirical results show better performance of proposed approach on the basis of *richer semantics* as compared to baseline approach. Even though our approach is quite simple, nonetheless it reveals interesting information about different conference mining tasks.

Possible future direction of this work is use of authors' information in addition to already used information for discovering research community. As we think, the research community discovered from CL will be more precise than that of DL due to topics denseness.

Acknowledgements. The work is supported by the National Natural Science Foundation of China under Grant (90604025, 60703059), Chinese National Key Foundation Research and Development Plan under Grant (2007CB310803) and Higher Education Commission (HEC), Pakistan. We are thankful to Jie Tang and Jing Zhang for sharing their codes, valuable discussions and suggestions.

References

1. Andrieu, C., Freitas, N.D., Doucet, A., Jordan, M.: An Introduction to MCMC for Machine Learning. Journal of Machine Learning 50, 5–43 (2003)
2. Azzopardi, L., Girolami, M., van Risjbergen, K.: Investigating the Relationship between Language Model Perplexity and IR Precision-Recall Measures. In: Proc. of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, July 28-August 1 (2003)
3. Balabanovic, M., Shoham, Y.: Content-Based Collaborative Recommendation. Communications of the ACM, CACM (1997)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
5. Blei, D.M., Lafferty, J.: Dynamic Topic Models. In: Proc. of 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, USA, June 25-29 (2006)
6. Breese, J., Heckerman, D., Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In: Proc. of the International Conference on Uncertainty in Intelligence (UAI), pp. 43–52 (1998)
7. Deshpande, M., Karypis, G.: Item-based Top-n Recommendation Algorithms. ACM Transactions on Information Systems 22(1), 143–177 (2004)
8. DBLP Bibliography database, `http://www.informatik.uni-trier.de/~ley/db/`
9. Girvan, M., Newman, M.E.J.: Community Structure in Social and Biological Networks. In: Proc. of the National Academy of Sciences, USA, vol. 99, pp. 8271–8276 (2002)
10. Griffiths, T.L., Steyvers, M.: Finding scientific topics. In: Proc. of the National Academy of Sciences, pp. 5228–5235 (2004)
11. Hofmann, T.: Probabilistic Latent Semantic Analysis. In: Proc. of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, July 30-August 1 (1999)
12. Kernighan, B.W., Lin, S.: An Efficient Heuristic Procedure for Partitioning Graphs. Bell System Technical Journal 49, 291–307 (1970)
13. Linstead, E., Rigor, P., Bajracharya, S., Lopes, C., Baldi, P.: Mining Eclipse Developer Contributions via Author-Topic Models. In: 29th International Conference on Software Engineering Workshops, ICSEW (2007)

14. Ley, M.: The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In: Proc. of the International Symposium on String Processing and Information Retrieval (SPIRE), Lisbon, Portugal, September 11-13, 2002, pp. 1–10 (2002)
15. McCallum, A., Nigam, K., Ungar, L.H.: Efficient Clustering of High-dimensional Data Sets with Application to Reference Matching. In: Proc. of the 6th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, August 20-23, 2000, pp. 169–178 (2000)
16. Popescul, A., Flake, G.W., Lawrence, S., et al.: Clustering and Identifying Temporal Trends in Document Databases. In: IEEE Advances in Digital Libraries (ADL), pp. 173–182 (2000)
17. Pothen, A., Simon, H., Liou, K.P.: Partitioning Sparse Matrices with Eigenvectors of Graphs. SIAM Journal on Matrix Analysis and Applications 11, 430–452 (1990)
18. Radicchi, F., Castellano, C., Cecconi, F., et al.: Dening and Identifying Communities in Networks. In: Proc. of the National Academy of Sciences, USA (2004)
19. Rosen-Zvi, M., Griffiths, T., Steyvers, M.: Smyth. P.: The Author-Topic Model for Authors and Documents. In: Proc. of the 20th International Conference on Uncertainty in Artificial Intelligence (UAI), Banff, Canada, July 7-11 (2004)
20. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: Extraction and Mining of Academic Social Networks. In: Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, USA, August 24-27 (2008)
21. Tyler, J.R., Wilkinson, D.M., Huberman, B.A.: Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. In: Proc. of the International Conference on Communities and Technologies, pp. 81–96 (2003)
22. Wang, X., McCallum, A.: Topics over time: A non-markov continuous-time model of topical trends. In: Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, August 20-23 (2006)
23. Wang, J.-L., Xu, C., Li, G., Dai, Z., Luo, G.: Understanding Research Field Evolving and Trend with Dynamic Bayesian Networks. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 320–331. Springer, Heidelberg (2007)
24. 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)
25. Zhang, J., Tang, J., Liang, B., et al.: Recommendation over a Heterogeneous Social Network. In: Proc. of the 9th International Conference on Web-Age Information Management (WAIM), ZhangJiaJie, China, July 20-22 (2008)