Appl Intell
DOI 10.1007/s10489-011-0302-3

Group topic modeling for academic knowledge discovery

Ali Daud · Faqir Muhammad

Abstract Conference mining and expert finding are useful academic knowledge discovery problems from an academic recommendation point of view. Group level (GL) topic modeling can provide us with richer text semantics and relationships, which results in denser topics. And denser topics are more useful for academic discovery issues in contrast to Element level (EL) or Document level (DL) topic modeling, which produces sparser topics. Previous methods performed academic knowledge discovery by using network connectivity (only links not text of documents), keywords-based matching (no semantics) or by using semantics-based intrinsic structure of the words presented between documents (semantics at DL), while ignoring semantics-based intrinsic structure of the words and relationships between conferences (semantics at GL). In this paper, we consider semantics-based intrinsic structure of words and relationships presented in conferences (richer text semantics and relationships) by modeling from GL. We propose group topic modeling methods based on Latent Dirichlet Allocation (LDA). Detailed empirical evaluation shows that our proposed GL methods significantly outperformed DL methods for conference mining and expert finding problems.

A. Daud (✉) · F. Muhammad
Department of Computer Science, Sector H-10, International Islamic University, Islamabad 44000, Pakistan
e-mail: ali.daud@iiu.edu.pk

F. Muhammad
e-mail: aioufsd@yahoo.com

F. Muhammad
Department of Business Administration, Sector E-9, Air University, Islamabad 44000, Pakistan

Published online: 11 June 2011

Keywords Denser topics · Conference mining · Unsupervised expert finding · Group topic modeling · Digital libraries

1 Introduction

Web is a great source of data and information convertible to knowledge. Many social networks have emerged due to the interactions of people on the web. We are certainly overwhelmed by the entities and their hidden relationships in these social networks. Automatic acquirement of useful information from text content has been a challenging problem, when most of the information is implicit within the entities (e.g. documents, researchers, conferences or journals) and their relationships in academic social networks, which are our focus in this work. For example, conferences are organized every year about different topics and huge volume of scientific literature is collected in digital libraries, such as DBLP and Citeseer. The data collected in these libraries provides us with many challenging academic knowledge discovery problems, which has many applications useful from researchers' point of view. For example, a new researcher should have guidance in obtaining authoritative conferences of specific research area to do literature review, a group of researchers would like to know about complete list of conferences related to their research area for submitting papers, program committee members are interested in conducting joint conferences, finding experts on specific topics for fulfilling reviewing and program committee tasks in conferences and journals, etc.

Conference mining and experts finding are highly investigated knowledge discovery problems in academic social networks for making useful recommendations to researchers.

Several methods proposed for academic knowledge discovery or related problems can be categorized into three major frameworks (1) graph connectivity based approaches as a basis for representation and analysis of relationships [3, 6, 9, 11, 14, 15, 19, 20, 23, 26, 27] (2) keywords-based matching using language models by exploiting TF-IDF [26, 28], and (3) topic modeling based approaches which make use of latent topic layer between words and documents to capture the text semantics-based relationships between entities [21, 22].

The main problem with the graph connectivity and keywords-based matching methods was ignorance of the text semantics-based information; consequently topic modeling came to overcome by using latent topic layer to model text semantics-based information. Unfortunately, recent topic modeling (DL) approaches [21, 22] either ignored conferences or viewed conferences information just as a stamp (token), which became the reason for ignoring implicit semantics-based text structure presented between the conferences. We think implicit text semantics-based information presented between the conferences (GL) is very useful and important for mining conferences and finding experts.

In this paper, we will consider semantics-based text structure and relationships presented between the conferences explicitly. We generalized previous topic modeling approach [22] idea of mining conferences and finding experts from a document level "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 and relationships* because of many semantically related words and authors are present in all documents of a conference), as a matter of fact the areas of interests in conference are highly correlated and overlapped, as are the accepted papers. It can provide grouping of conferences in different groups on the basis of latent topics (semantically related probabilistic cluster of words) presented between the conferences or group. We propose a Latent Dirichlet Allocation (LDA) [4] based GL ConMin approach for conference mining and temporal expert topic approach (TET) for finding experts. Empirical results showed that GL based proposed methods clearly achieve better results than DL idea based methods for both academic knowledge discovery problems by capturing richer text semantics and relationships at group level resulting in denser topics. Solution provided by us produced quite intuitive and functional results.

The contributions of this work includes

(1) Formalization of the key conference mining issues
(2) Proposal of group topic modeling (ConMin) approach to deal with the issues by capturing *richer semantics with* experimental verification of the effectiveness of our approach on real-world large dataset
(3) To give notion of dense topics and demonstration of their positive impact on models performance in topic modeling domain
(4) Proposal of group topic modeling for unsupervised expert finding with proven effectiveness

To the best of our knowledge, we are the first to deal with the aforementioned academic knowledge discovery issues directly by proposing GL topic modeling approaches, which can produce dense topic as compared to sparse topics produced by DL topic modeling approaches.

The rest of the paper is organized as follows. In Sect. 2, we formalize the key conferences mining issues and expert finding problem followed by the typical topic modeling and its adoptions for solving both academic knowledge discovery problems. In Sect. 3, dataset, parameters settings, performance measures and baseline approach is given. Section 4 provides results and discussions with showing comparisons for both problems in detail and Sect. 5 brings this paper to the conclusions.

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. Conference level (CL) and group level (GL) is also used interchangeably.

2 Knowledge discovery in academic social networks

In this section, before describing our ConMin and TET approaches for academic knowledge discovery, we will first formulize conference mining tasks and expert finding problem, describe state-of-the-art topic model LDA [4], followed by modeling of conferences with authors based topics (ACT1) [22].

2.1 Problem formulization

Conference mining through their accepted papers by considering group level text semantics and relationships are intuitive. Each conference accepts many papers every year related to some sort of overlapped areas of research. To our interest, each publication contains title which covers most of the highly related sub research areas. Conferences with their accepted papers titles on the basis of their latent topics can be mined in a better way as compared to documents or authors based topics. We only used paper titles as experiments have shown that using the whole text of papers and using only titles of papers do not affect the performance of methods much but on the other hand reduce the time complexity to a great extent. 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 level (Constituent-Document).

(1) Topics based Ranking of Conferences: 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 Conference 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.

Temporal expert finding addresses the task of finding people who are experts in some domain for different time periods (e.g. years in this work). Expert finding became one of the biggest challenges in enterprises and time is important as one expert cannot be expert for his whole life. We put emphasis on temporal expert finding rather than general expert finding so as to support questions like "Who are the experts on topic Z for year Y? Instead of just who are the experts on topic Z?" A submitted query is denoted by q and an expert is denoted by m. In general semantics-based temporal expert finding process, the main task is to probabilistically rank discovered experts for a given query for different years, where a query is usually comprised of several words or tokens and token is referred to as a collection of words as one term such as Data Mining. Intuition behind exploiting conferences richer text semantics and relationships is based on the thinking that high level conferences usually have more semantically related words and participating authors publishing in these are usually experts as compared to authors participating in low level conferences.

2.2 Latent Dirichlet allocation (LDA)

Before introducing LDA, we describe the limitations of keywords and traditional clustering methods. Keywords based modeling uses exact word matching for finding related entities, example of which is state-of-the-art vector space model (VSM) [28]. Clustering provides a good way to group similar documents for automatic extraction of topics from text [17, 18] based on similar contents. The problem with keywords based matching is ignorance of semantics or in other words synonymy and polysemy and traditional clustering is inherently limited by the fact that each document is only associated with one cluster, which motivated latent topic layer based topic modeling. Topic models are soft clustering representation techniques, which can capture text semantics and

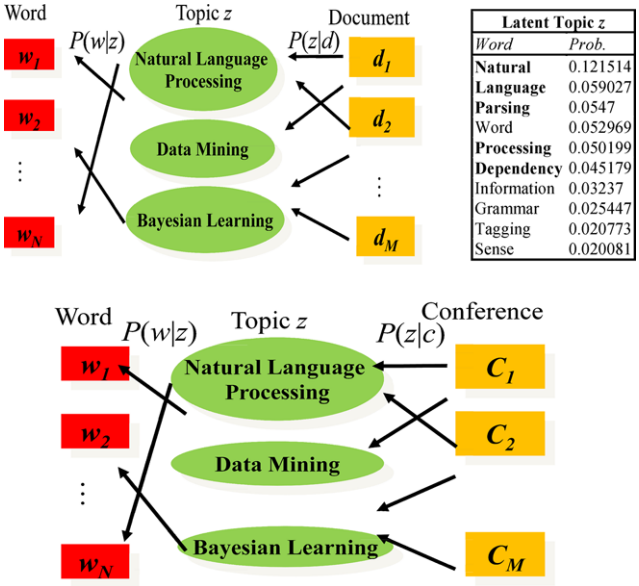


Fig. 1 Document level (DL) *up* and conference level (CL) *down*, topic modeling

allow documents composed of multiple topics to relate to more than one cluster on the basis of latent topics.

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 $\mathbf{w}_d$. This topic layer is proved very useful for capturing semantics-based relationships by considering synonymy and polysemy of words. Basically, 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.

Figure 1 provides pictorial representation of typical topic modeling, in which latent topic layer is used between words and documents to match documents with the queries. We explain it with the help of an information retrieval example. Suppose a user enters a query natural language processing for which following two papers are retrieved. First paper title contains the query words natural language processing so found related to the query, while second paper title includes dependency parsing not included in the user query words even then it is found related to a query because of semantic similarity of natural language processing and dependency parsing words in a topic "Natural Language Processing" whose top ten words with their assigned probabilities are shown in Fig. 1 top. Figure 1 bottom provides pictorial representation of generalized topic modeling, in which latent topic layer is used between words and conferences for answering queries with richer text semantics or denser topics.

– Paper1: A Maximum Entropy Approach to *Natural Language Processing*
– Paper2: A Pipeline Framework for *Dependency Parsing*

LDA [4] is a state-of-the-art topic modeling approach which makes use of latent topic layer to capture semantics-based relationships between words. In its generative process first, for each document d in corpus, a multinomial distribution θ_d over topics is randomly sampled from a Dirichlet distribution with parameter α. Second, for each word w in a 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 to produce documents. Simply the generating probability of word w from document D for LDA is given as:

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

2.3 Modeling conferences with authors topics (ACT1 (DL))

LDA has been extended for solving different research problems. Recently, it is extended to discover topically related conferences indirectly by using topics of documents generated by authors in ACT1 model [22]. The basic idea of topic modeling that words and documents can be modeled by considering latent topics and later modeling words and authors of documents [21] became the intuition of modeling the words, authors and conferences through latent topics. The generative process of ACT1 is based on the idea that initially, authors think of writing a research paper on a topic and correspondingly select the conference to submit it.

Technically in ACT1, 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:

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

The generative process is as follows:

1. For each author $r = 1,\ldots,K$ of document d
   Choose θ_r from Dirichlet (α)
2. For each topic $z = 1,\ldots,T$
   Choose Φ_z from Dirichlet (β)
   Choose Ψ_z from Dirichlet (γ)
3. For each word $w = 1,\ldots,N_c$ of document d
   Choose an author r uniformly from all authors $\mathbf{a}_d$
   Choose a topic z from multinomial (θ_r) conditioned on r
   Choose a word w from multinomial (Φ_z) conditioned on z
   Choose a conference stamp c associated with word w from multinomial (Ψ_z) conditioned on z

2.4 Modeling conferences with topics (ConMin (GL))

The basic idea of topic modeling that words and documents can be modeled by considering latent topics became the intuition for modeling the words and conferences directly through latent topics. The intuition of our proposed ConMin approach is based on the fact that for finding topically related conferences, conference relationships and temporal topic trends, conferences based latent topics are more important as compared to authors based latent topics [22]. Authors sometimes have diverse kind of research interests and they are also publishing in many diverse conferences and journals which may result in generating very sparse topics in this way. Sparse topics mean high perplexity for that specific approach which usually results in vague cluster of probabilistically related words (latent topics). Consequently, we generalize this idea from DL [4] to GL by considering documents as sub-entities of a conference to explore conferences based topics, which may be dense based on the previous discussion.

In our approach a conference is viewed as a composition of the words of all its accepted publications. Symbolically, for a conference c we can write it as: $C = \{d_1 + d_2 + d_3 + \ldots + d_i\}$, where d_i is one document in a conference.

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 for a conference contained in super-document, a topic z is chosen from this topic distribution. Finally, the word w is generated by randomly choosing 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:

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

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

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

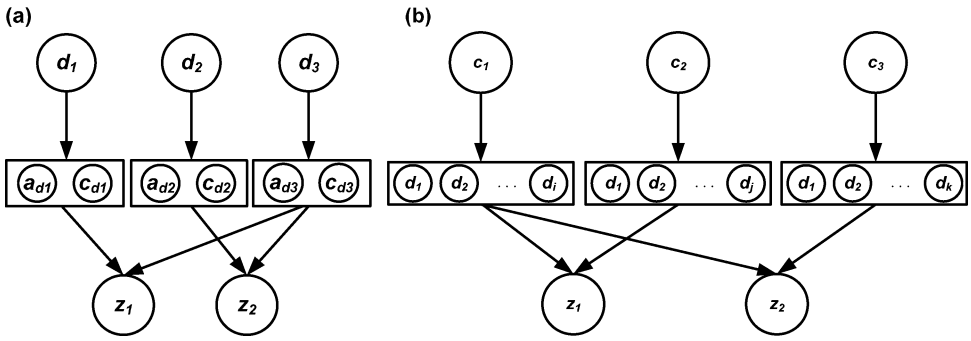


Fig. 2 Conference modeling (**a**) ACT1 (DL) and (**b**) ConMin (CL) approaches

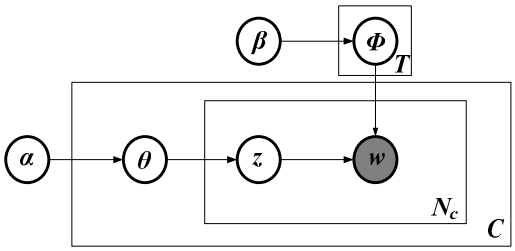


Fig. 3 ConMin (generalized smoothed LDA)

where $z_i = j$ represents the assignments of the ith word in a conference to a topic j. $\mathbf{z}_{-i}$ represents all topic assignments excluding the ith 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. "$\cdot$" 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 $W \times Z$ (words by topic) and $Z \times C$ (topic by conference) count matrices. From these count matrices, topic-word distribution Φ and conference-topic distribution ϑ can be calculated as given in (5) and (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

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

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

2.5 Modeling experts with topics (TET (GL))

We investigate the problem of temporal expert finding in an insupervised way by simultaneously modeling conferences influence or group level influence, another application of our GL idea for conference mining tasks. From unsupervised means we do not need to know his exact number of publications and their citations, his academic activities such as program committee member, editorial board member etc. And usually to collect all this information about all researchers is cumbersome. A support vector machine based methods is proposed for identifying the authors of documents [30] and continuum of general to specific interests of a user is extracted to provide more robust personalization [29], which are sister problems of expert finding.

We proposed group time topic modeling approach Temporal Expert Topic (TET), which can provide ranking of experts in different groups in an unsupervised way. It is generalized from previous topic model ACT1 [22] form a single document "sub-group" (*no conferences influence or document level*) to all publications of the conference "Group" (*conferences influence or group level*). We treat ACT1 as baseline for expert finding task from document level to group level.

The intuition behind considering conferences as a Group is explained with the help of an example in Fig. 4. A document denoted as a subgroup here, usually has a few semantically related words (as total words in title are only "8") and authors (as total authors are only "2") to a topic shown in Fig. 4, while a conference denoted as a "Group" here, usually there are many related papers to a topic; as a result a Group usually has many semantically related words (as total words are as high as "439") and authors (as total authors is as high as "95") to a topic as shown in Fig. 4. Subgroup is a subset of a group as highlighted in Fig. 4; consequently semantic-based information and relationships are richer in a group as compared to a subgroup, which is referred to as "Conferences Influence" in our work and main contribution of this work. Our thinking is supported by the facts that (1) in highly ranked events usually papers of experts or potential experts of different fields are accepted, therefore event based relationships are highly influential which reminds us a famous saying "A man is known by the company he keeps" and (2) accepted papers in highly ranked events

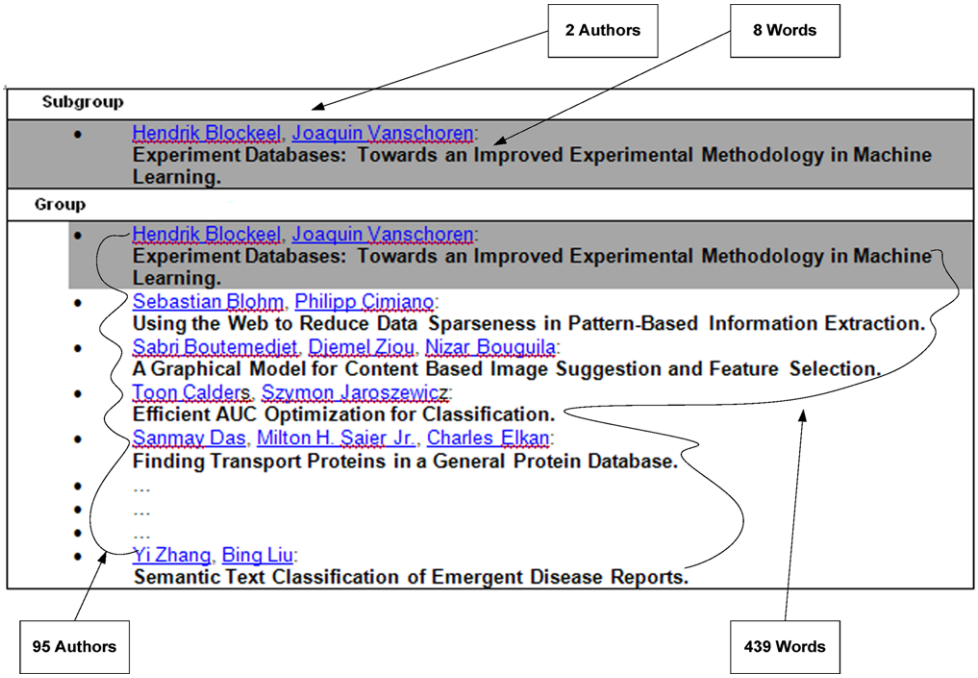


Fig. 4 A group illustration for accepted papers by ECML/PKDD-2007

are very carefully judged for relevance to the event research areas on call for papers page, therefore papers have more semantically related words and authors, which can result in higher ranking of their authors because of conferences influence.

Non-Generalized Topic Modeling approach ACT1 [22] uses conferences information just as a token, which results in not capturing the conferences influence and time information is also not modeled simultaneously in it. Consequently, we propose generalized time topic modeling approach named Temporal-Expert-Topic (TET), which can utilize both conferences influence and time information, simultaneously.

In TET, each author from a set of K authors of a conference is considered responsible for generating some latent topics of a conference and in turn these topics generate the words and time stamps for that conference. Formally, each author from a set of K authors of an event c is associated with a multinomial distribution θ_r over topics and each topic is associated with a multinomial distribution Φ_z over words and multinomial distribution Ψ_z with a year stamp for each word of an event for that topic. So, θ_r, Φ_z and Ψ_z have a symmetric Dirichlet prior with hyper parameters α, β and γ, respectively. The generating probability of the word w with year y for author r of event c is given as:

$$P(w, y|r, c, \emptyset, \Psi, \theta) = \sum_{z=1}^{T} P(w|z, \emptyset_z) P(y|z, \Psi_z) P(z|r, \theta_r) \quad (7)$$

3 Experiments

3.1 Dataset

DBLP online database [10, 16] is a huge source of research publications and related information which is very useful from academic social network analysis point of view. Five year publication dataset of conferences is downloaded from the DBLP by only considering conferences for which data was available for years 2003–2007. We selected conferences in this way to make sure that these conferences are regular one and are being organized every year. Totally, we extracted 90,124 publications for 261 conferences and combined them into a super-document separately for each conference. We then preprocessed corpus by using typical preprocessing procedures adopted for text mining by (a) removing stop-words standard list, punctuations and numbers from the words (b) down-casing the obtained words for proper string matching, and (c) removing words that appear less than three times in the corpus to simply ignore words which are seldom used by authors to name their method such as ConMin in this paper, which may not have any meaning if it is not written in full as Conference mining. This led to a vocabulary size of $V = 10{,}902$ and a total of 571,439 words in the corpus. Another reason of selecting alike conferences for similar years is to precisely analyze the conferences temporal trends. Figure 5 shows quite smooth yearly data distribution for number of publications in conferences.

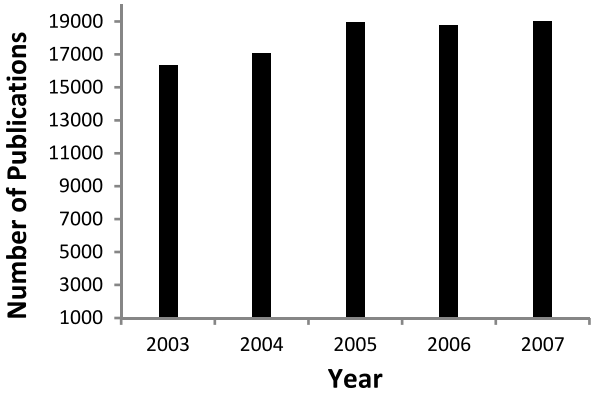


Fig. 5 Yearly conferences publications

3.2 Parameter settings

Parameter estimation for text analysis can be performed by using different methods. One can estimate the optimal values of hyper-parameters α and β (Fig. 3) by using Expectation Maximization (EM) method [13] or Gibbs sampling algorithm [12] by maximizing the likelihood. 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 ConMin 200 topics Z, the hyper-parameters α and β were set at $50/Z$ and .01, respectively [7]. There is no hard and fast rule to set the number of topics although perplexity is considered as one of the matrix useful both for checking model performance evaluation and setting number of topics. We calculated the perplexity for number of topics from 2, 5, 10, 20, 40, . . . , 300 and number of topics Z were fixed at 200 based on measured perplexity [2] on 20% held out test dataset plus on the basis of human judgment of meaningful topics. 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 for ConMin.

3.3 Performance measures

For conference mining issues, performance evaluation is performed both qualitatively and quantitatively. 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 and usually used to evaluate the performance of clustering approaches. Secondly, we used average Symmetric KL (sKL) divergence [21] 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 as it explains that the boundaries of topics have less overlaps or topics are more refined clusters of probabilistic words in terms of clustering. Following equations are used for calculating entropy and sKL divergence. In (8), we used topic-word distribution matrix for all words of each topic to calculate intra-topic entropy and then calculated the average entropy for all topics. In (9), we used conference-topic distribution matrix for conferences calculating the inter-topic difference between conferences i and j

$$\textit{Entropy of (Topic)} = -\sum_{z} P(z)\log_2[P(z)] \tag{8}$$

$$\text{sKL}(i,j) = \sum_{z=1}^{T}\left[\theta_{iz}\log\frac{\theta_{iz}}{\theta_{jz}} + \theta_{jz}\log\frac{\theta_{jz}}{\theta_{iz}}\right] \tag{9}$$

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 employ 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 "Digital Libraries" topic it is JCDL) 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 in Table 1.

For expert finding, we provide comprehensive (DBLP data Statistics) based comparison [10] in Table 5, for 150 topics for our proposed and baseline approach. In it, we show how our proposed approach produced more precise results because of (1) top ten experts in list published more in the World Level (World Class) conferences, (2) from top 3 conferences for each expert most of the time at least one of them is world level and (3) number of papers published by top ten experts list for the topics is also greater.

3.4 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 [22]. We used the same machine used for proposed ConMin approach; run time per each chain for ACT1 was 3.00 hours almost double than ConMin, which was 1.26 hours. It shows that ConMin is

Table 1 An illustration of 7 discovered topics (top ConMin approach, bottom ACT1 approach)

Topic 117 (ConMin) "XML Databases"		Topic 164 (ConMin) "Semantic Web"		Topic 63 (ConMin) "Information Retrieval"		Topic 138 (ConMin) "Digital Libraries"		Topic 190 (ConMin) "Data Mining"		Topic 28 (ConMin) "Bayesian Networks"		Topic 0 (ConMin) "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

Table 1 (*Continued*)

Topic 109 (ACT1) "XML Databases"		Topic 154 (ACT1) "Semantic Web"		Topic 163 (ACT1) "Information Retrieval"		Topic 122 (ACT1) "Digital Libraries"		Topic 10 (ACT1) "Data Mining"		Topic 82 (ACT1) "Bayesian Networks"		Topic 37 (ACT1) "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

also better in terms of time complexity for mining conferences.

4 Results and discussions

4.1 Conference mining

4.1.1 Comparisons

Topically related conferences based comparison The effect of topic sparseness on the approach 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 based semantic relationships between conferences. Table 1 illustrates 7 different topics out of 200, discovered from the 1000th iteration of a particular Gibbs sampler run. Each topic is shown with the top 10 words and conferences and titles are our interpretation of the topics. The words associated with each topic for our proposed approach are strongly semantically related (less sparse) than that of baselines, 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 Fig. 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).

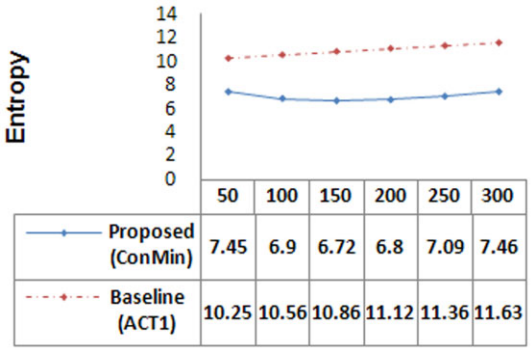
For example, "Web Search" topics related top ten conferences list for proposed approach begins with WWW, LA-WEB, ..., ASWC with corresponding probabilities from 0.23, 0.21, ..., 0.061, while for same topic top ten conference list for baseline approach begins with WWW, CIKM, ..., ADBIS with corresponding probabilities from 0.98, 0.0013, ..., 0.000034. One can clearly see that the corresponding probabilities for baseline approach are highly skewed and WWW conference has very high probability 0.98, while other "Web Search" related conferences are assigned very low probabilities, which is against the real world situation. Similar kind of skewness problem is observed in all topically related conferences in Table 1 for baseline approach.

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 as they should have to found for "Data Mining" topic first and then for some other topics like "Web Search". 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 or top tier conferences in that area of research, but rather are the conferences that tend to produce most semantically related 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 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.

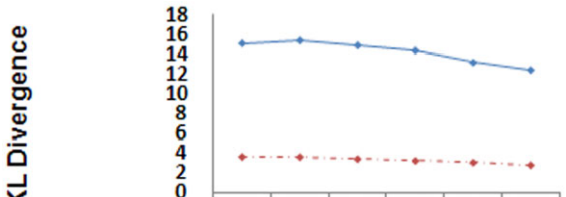
Proposed approach discovers several other topics related to data mining such as neural networks, multi-agent systems and pattern matching, also other topics that span the full range of areas encompassed in the dataset. A fraction of non-research topics, perhaps 10–15%, are also discovered that are not directly related to a specific area of research, as the words present in those topics were actually used as a glue between scientific terms.

Entropy based comparison In addition to qualitative comparison between ConMin and ACT1, we also provide quan-



	50	100	150	200	250	300
Proposed (ConMin)	7.45	6.9	6.72	6.8	7.09	7.46
Baseline (ACT1)	10.25	10.56	10.86	11.12	11.36	11.63

Fig. 6 Average entropy curve as a function of different number of topics, lower is better



	50	100	150	200	250	300
Proposed (ConMin)	15.03	15.31	14.83	14.32	13.1	12.35
Baseline (ACT1)	3.56	3.54	3.34	3.17	2.98	2.72

Fig. 7 Average sKL divergence curve as a function of different number of topics, higher is better

titative comparison to explain the effect of topics sparseness on the performance of approach. Figure 6 shows the average entropy of topic-word distribution for all topics measured by using (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.

Symmetric KL divergence based comparison Figure 7 shows the average distance of topic-word distribution between all pairs of the topics measured by using (9). 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 Figs. 6 and 7 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.

Error rate based comparison 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.

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 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 our proposed approach.

4.1.2 Applications of ConMin

Conferences correlations ConMin and ACT1 both approaches can be used for automatic correlation discovery [21] 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 (9) for topics distribution conditioned on each of the conferences distribution.

We calculated the dissimilarity between the conferences by using (9), 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 3 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

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						

Table 3 *sKL* divergence for pairs of conferences of ConMin and ACT1

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

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 proposed and baseline 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 on ACT1 in capturing semantics-based correlations between conferences.

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 (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 4 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.

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 [4]:

$$perplexity(C_{test}) = \exp\left[-\frac{\log p(W_c)}{N_c}\right] \tag{10}$$

Figure 8 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.

Conferences temporal topic trends Temporal topic trends of computer science were discovered in Citeseer documents [18, 21] by utilizing clustering and semantics-based text information. Recently, Dynamic Topic model and Topics over Time [5, 24] 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 [25]. 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.

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

Table 4 An illustration of top five predicted topics for SIGIR and KDD conferences

Topic words	Title	Probability
SIGIR		
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		
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

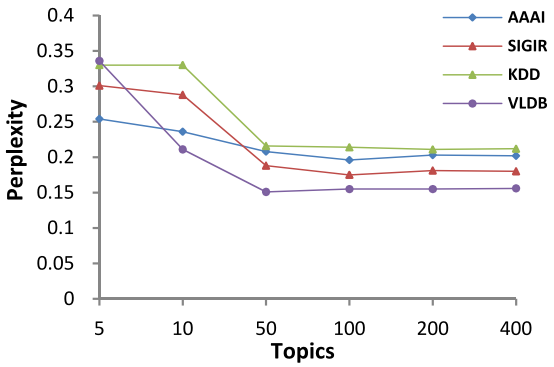


Fig. 8 Measured perplexity for new conferences

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 9 shows the results of plotting topics for SIGIR and KDD, where each topic is indicated in the legend with the five most probable words.

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 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.

4.2 Unsupervised expert finding

Based on the GL text semantics and relationships between authors idea we show temporal expert finding comparison between our proposed and baseline approach for DM topic related top 10 experts. Here top ten experts related to a topic for a year 2003 with the top 3 conferences are shown in which they published and number of papers they have published in that year. Results are just based on the subset data collected from the DBLP database for showing conferences influence or GL influence and cannot be used for exact comparison between the authors. For this purpose we have not provided the names of authors instead we provide their id numbers given by us without any specific order.

To show the dominancy of our proposed approach over the baseline approach which does not consider group level text semantics and authors relationships, we provide comparison of all years for DM topic by using DBLP database [10] provided statistics for each expert. For this purpose we divided conferences into two main categories, World Level "WL" (Considered better than normal level due to their high

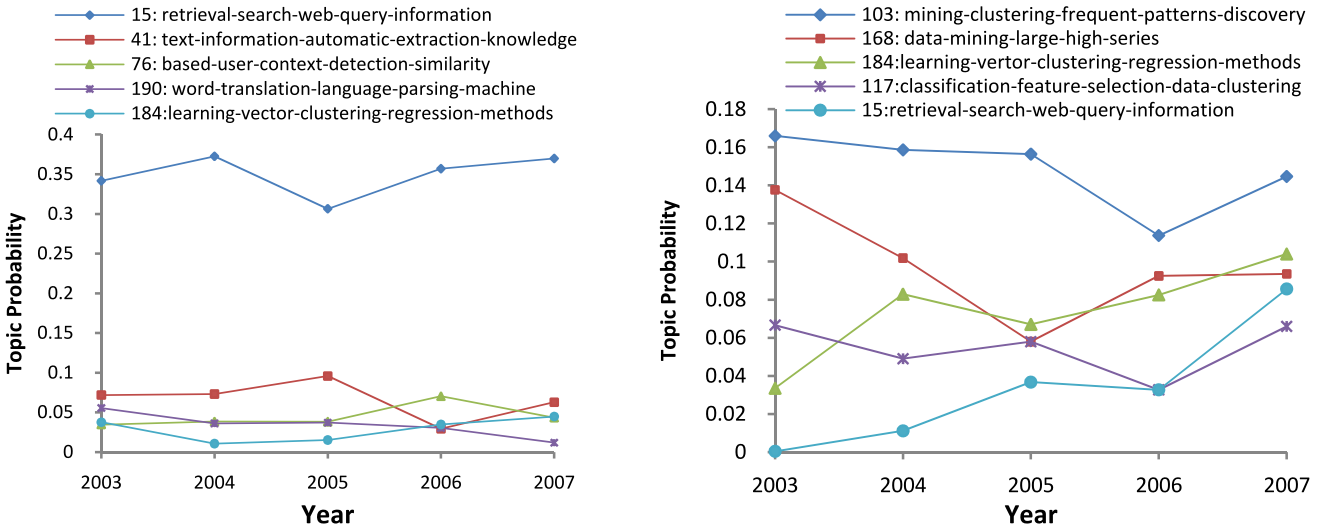


Fig. 9 Temporal topic trends of conferences

Table 5 Temporal expert finding comparison between our proposed and baseline approach for DM topic related top 10 experts

Experts	2003 data mining (TET) Top 3 conferences	TP	Experts	2003 data mining (ACT1) Top 3 conferences	TP
2628	WL (ICDE), NL (ISCAS, ICDM)	33	4477	WL (ICDE, KDD, SIGMOD)	19
5135	NL (BIBE, DAWAK, GRC)	9	2681	NL (ICDM, IDEAS, MDM/KDD)	10
5119	WL (ICDE, SIGMOD), NL (DASFAA)	13	5018	NL (ICDM, ADMA, APIN)	6
4477	WL (ICDE, KDD, SIGMOD)	19	2231	NL (AAI, ADC, AI)	5
2630	WL (KDD), NL (ICDM, IPDPS)	11	1660	WL (ICDE, SIGMOD), NL (ICDM)	12
118	WL (SIGIR), NL (ICDM, CIKM)	12	2630	WL (KDD), NL (ICDM, IPDPS)	11
4786	WL (KDD), NL (ICDM, SDM)	14	8642	NL (ICEIS, ICWI, IKE)	9
1659	WL (KDD), NL (ICDM, SDM)	6	323	NL (CIKM, PAKDD, ICDCS)	19
5014	WL (SIGIR, WWW), NL (ICDM)	8	5325	WL (KDD), NL (ICTAI, DASFAA)	9
5017	NL (ICEIS, SAC, IRI)	14	8737	NL (CIKM, PAKDD, IDEAS)	7

class) and Normal Level "NL" or others to evaluate the performance of approaches in terms of considering and not considering conferences influence. Here for DM topic *KDD, ICDE, SIGMOD, VLDB, WWW*, and *SIGIR* are considered as WL conferences (on the basis of expert opinions and impact scores on Citeseer *(http://citeseer.ist.psu.edu/)* and others are considered as NL conferences. We just made two categories for simplicity and to show generalization time topic modeling effectiveness over the baselines one can make as many categories as he/she like. Top three conferences for each author are selected from DBLP data statistics [10] and categorized them as WL (bold font in Top 3 Conferences column) and NL (normal font in Top 3 Conferences column) in Table 5. Total Papers (TP) column shows number of papers published in a given year by the expert in all conferences.

In Table 6, WL means World Class conference, OneWL means at least one conference is WL in top 3 conferences related to an expert and TP means total number of papers for top ten topically related authors to a topic. We can see in Table 6 firstly, for year 2003 of TET from top ten experts 12 times papers are published in WL conferences with total number of 139 papers and for year 2003 of ACT1 [22] from top ten experts 7 times papers are published in WL conferences with total number of 107 papers. 12 times WL for TET is greater than 7 times WL for ACT1, which shows that authors found by our proposed approach have comparatively more expertise on topic as compared to baseline.

Secondly, 8 experts from top ten shown for TET at least have OneWL conference related to an expert in top 3 conferences and 4 experts from top ten for ACT1 at least have OneWL conference related to an expert in top 3 conferences. 8 experts OneWL for TET are greater than 4 experts

Table 6 Summary of Table 5

Year	WL (ACT1)	WL (TET)
2003	7	12
2004	11	11
2005	12	12
2006	10	13
2007	10	12
Average	10	12
Year	OneWL (ACT1)	OneWL (TET)
2003	4	8
2004	6	7
2005	6	8
2006	6	9
2007	6	8
Average	5.6	8
Year	TP (ACT1)	TP (TET)
2003	107	139
2004	204	209
2005	201	248
2006	293	298
2007	219	289
Average	204.8	236.6

for ACT1. Thirdly, 139 TP for TET are greater than 107 TP for ACT1. It clearly shows that experts found by TET approach are better, as they published more in WL conferences, more experts in the top ten lists has published at least in one OneWL and experts published more papers as compared to document level baseline approach ACT1. The above situation is also true for years 2004, 2005, 2006 and 2007.

Thirdly, Table 6 shows that the average number of times experts publishing in WL 12 for TET is greater than WL 10 of ACT1, average number of experts publishing at least in one world class conference average OneWL is 8 for TET that is greater than average OneWL 5.6 for ACT1, which supports our hypothesis that our approach can discover more precise experts who published more in WL conferences than experts discovered by document level approach.

One can say that if someone is expert of some area of research he should have at least one world class conference among his/her top three publishing conferences. Additionally, average number of papers for TET approach for top ten experts is 236.6 which are greater than average number of papers for ACT1 approach 204.8, which shows the proposed approach acquiring more accurate results.

The results presented in Table 6 show that TET outperformed ACT1 due to its ability to simultaneously capturing conferences influence with time information.

5 Conclusions

This study deals with two important problems of academic knowledge discovery through capturing rich text semantics-based structure of words and relationships present between conferences at group level. We conclude that our generalization from DL to GL is significant; as proposed GL approach's discovered conferences and their correlations (can also be applied to journals datasets such as HEP or OHSUMED) related to specific knowledge domains are better than baseline approach due to producing dense topics. We studied the effect of generalization on topics denseness and concluded that sparser topics will result in poor performance of the approach. We have also shown the effectiveness of conferences influence (text semantics and relationships at GL) for expert finding problem. Even though our GL approaches are quite simple, nonetheless they reveal practical importance over DL approach for different conference mining tasks and unsupervised expert finding problem.

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

The material in this paper was presented in part at 19th International European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML/PKDD 2009) [8].

References

1. Andrieu C, Freitas ND, Doucet A, Jordan M (2003) An introduction to MCMC for machine learning. J Mach Learn 50:5–43
2. Azzopardi L, Girolami M, van Risjbergen K (2003) 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 (1997) Content-based collaborative recommendation. Commun ACM
4. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
5. Blei DM, Lafferty J (2006) 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 (1998) Empirical analysis of predictive algorithms for collaborative filtering. In: Proc of the international conference on uncertainty in intelligence (UAI), pp 43–52
7. Daud A, Li J, Zhu L, Muhammad F (2010) Knowledge discovery through directed probabilistic topic models a survey. J Front Comput Sci China 4(2):280–301
8. Daud A, Li J, Zhu L, Muhammad F (2009) Conference mining via generalized topic modeling. In: Buntine W et al (ed) Proc of European conference on machine learning and principles and practices of knowledge discovery in databases (ECML PKDD), Part I. LNAI, vol 5781, pp 244–259

9. Deshpande M, Karypis G (2004) Item-based top-n recommendation algorithms. ACM Trans Inf Sys 22(1):143–177
10. DBLP bibliography database. http://www.informatik.uni-trier.de/~ley/db/
11. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. In: Proc of the national academy of sciences, USA, vol 99, pp 8271–8276
12. Griffiths TL, Steyvers M (2004) Finding scientific topics. In: Proc of the national academy of sciences, pp 5228–5235
13. Hofmann T (1999) Probabilistic latent semantic analysis. In: Proc of the 15th annual conference on uncertainty in artificial intelligence (UAI), Stockholm, Sweden, July 30–August 1, 1999
14. Kernighan BW, Lin S (1970) An efficient heuristic procedure for partitioning graphs. Bell Syst Tech J 49:291–307
15. Linstead E, Rigor P, Bajracharya S, Lopes C, Baldi P (2007) Mining eclipse developer contributions via author-topic models. In: 29th international conference on software engineering workshops (ICSEW)
16. Ley M (2002) 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
17. McCallum A, Nigam K, Ungar LH (2000) 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
18. Popescul A, Flake GW, Lawrence S et al. (2000) Clustering and identifying temporal trends in document databases. IEEE Adv Digit Libr 173–182
19. Pothen A, Simon H, Liou KP (1990) Partitioning sparse matrices with eigenvectors of graphs. SIAM J Matrix Anal Appl 11:430–452
20. Radicchi F, Castellano C, Cecconi F et al (2004) Dening and identifying communities in networks. In: Proc of the national academy of sciences, USA
21. Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) 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
22. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z (2008) 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
23. Tyler JR, Wilkinson DM, Huberman BA (2003) Email as spectroscopy: automated discovery of community structure within organizations. In: Proc of the international conference on communities and technologies, pp 81–96
24. Wang X, McCallum A (2006) 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
25. Wang J, Xu C, Li G, Dai Z, Luo G (2007) Understanding research field evolving and trend with dynamic Bayesian networks. In: Proc of the PAKDD
26. Zaiane OR, Chen J, Goebel R (2007) DBconnect: mining research community on DBLP data. In: Joint 9th WEBKDD and 1st SNA-KDD workshop, San Jose, California, USA, August 12, 2007
27. Zhang J, Tang J, Liang B et al (2008) Recommendation over a heterogeneous social network. In: Proc of the 9th international conference on web-age information management (WAIM), ZhangJia-Jie, China, July 20–22, 2008
28. Zhai C, Lafferty J (2001) A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proc of the 24th ACM SIGIR international conference on information retrieval, pp 334–342
29. Kim HR, Chan PK (2008) Learning implicit user interest hierarchy for context in personalization. J Appl Intell 28:153–166
30. Diederich J (2003) Authorship attribution with support vector machines. J Appl Intell 19:109–123

Ali Daud is working as Assistant Professor in the Department of Computer Science at International Islamic University, Islamabad. He obtained his PhD degree from Tsinghua University in 2010. He is head of databases group and published about 14 papers in international journals. He has taken part in many projects and PI of a project funded by higher education commission, Pakistan. His current research interests includes: text mining, social networks analysis and applications of probabilistic topic models.

Faqir Muhammad is a full professor at Air University, Islamabad. He obtained his PhD degree from University of Glasgow in 1987. He has published about 50 papers in international journals and conferences. His main research interests include statistical modeling, multivariate analysis, text mining and knowledge discovery.