Knowledge-Based Systems xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Temporal expert finding through generalized time topic modeling

Ali Daud [a,*], Juanzi Li [a], Lizhu Zhou [a], Faqir Muhammad [b]

[a] Department of Computer Science and Technology, 1-308, FIT Building, Tsinghua University, Beijing 100084, China
[b] Department of Mathematics and Statistics, Allama Iqbal Open University, Sector H-8, Islamabad 44000, Pakistan

ARTICLE INFO

Article history:
Received 5 May 2009
Received in revised form 6 April 2010
Accepted 10 April 2010
Available online xxxx

Keywords:
Temporal expert finding
Conferences influence
Generalized time topic modeling
Unsupervised learning

ABSTRACT

This paper addresses the problem of semantics-based temporal expert finding, which means identifying a person with given expertise for different time periods. For example, many real world applications like reviewer matching for papers and finding hot topics in newswire articles need to consider time dynamics. Intuitively there will be different reviewers and reporters for different topics during different time periods. Traditional approaches used graph-based link structure by using keywords based matching and ignored semantic information, while topic modeling considered semantics-based information without conferences influence (richer text semantics and relationships between authors) and time information simultaneously. Consequently they result in not finding appropriate experts for different time periods. We propose a novel Temporal-Expert-Topic (TET) approach based on Semantics and Temporal Information based Expert Search (STMS) for temporal expert finding, which simultaneously models conferences influence and time information. Consequently, topics (semantically related probabilistic clusters of words) occurrence and correlations change over time, while the meaning of a particular topic almost remains unchanged. By using Bayes Theorem we can obtain topically related experts for different time periods and show how experts' interests and relationships change over time. Experimental results on scientific literature dataset show that the proposed generalized time topic modeling approach significantly outperformed the non-generalized time topic modeling approaches, due to simultaneously capturing conferences influence with time information.

1. Introduction

The World Wide Web is the biggest source of diverse type of information which keeps on changing with respect to time. Particularly, various specialized digital libraries stores enormous amounts of time series data. Automatic discovery of useful information while capturing its dynamisms from these libraries is an interesting and challenging issue discussed recently. With the advancement of information retrieval technologies from traditional document-level to object-level [34], expert finding problem has gained a lot of attention in the web-based research communities. The motivation is to find a person with topic relevant expertise to automatically fulfill different recommendation tasks by creating knowledge bases of researchers that would support the finding of appropriate collaborators for the project, choosing experts for consultation about research topics, matching reviewers with research papers, to finalizing program committee members and inviting keynote speakers for the conferences.

In the past, major frameworks used for expert finding can be divided into two main categories (1) graph connectivity based approaches [7,25,27] which make use of graph links on the basis of paper citations or co-authorships by using keywords based matching and (2) semantics-based topic modeling approaches which make use of latent topic layer to semantically capture relationships. Firstly, above mentioned frameworks based on network connectivity ignore the semantics-based information present in the documents. Unfortunately, topic modeling captures the semantics-based information present in the documents but ignored the conferences influence. Secondly, most of the existing topic modeling approaches for expert finding ignored simultaneous modeling of time information results in the exchangeability of topics problem, which means the inability of the topic model not to obtain similar topics for different years.

These days most of the datasets such as research papers, blogs and news do not have static co-occurrence patterns; they are instead highly dynamic. The data are collected over time and data patterns keep on changing, by showing rising or falling trends with respect to time, therefore it is not a realistic assumption to ignore time factor. Illustratively in the temporal expert finding problem (1) expert A could change his research interest, e.g. expert A mainly

* Corresponding author.
E-mail addresses: ali_msdb@hotmail.com (A. Daud), ljz@keg.cs.tsinghua.edu.cn (J. Li), dcszlz@tsinghua.edu.cn (L. Zhou), aioufsd@yahoo.com (F. Muhammad).

0950-7051/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2010.04.008

focused on natural language processing until 2004 and published a lot of papers about this topic; afterwards he switched his concentration to academic social networks analysis and did not published many papers. He can still be found as an expert in natural language processing topic in 2010 if the time factor were ignored, but he might not be an appropriate choice any longer and (2) new authors could be writing on similar topics with expert A and can push back his ranking for a specific time. Intuitively, ranked experts related to a specific topic and their relationships for each given year cannot be the same.

Here it is necessary to mention that exploitation of authors' interests [23] (who is writing on what topic without any discrimination between renowned and not-renowned publication events) and expert finding [1] (who is most skilled on what topic with the discrimination between renowned and not-renowned publication events) are notably two different knowledge discovery problems.

In this paper, we investigate the problem of temporal expert finding by simultaneously modeling conferences influence and time information. We proposed the generalized time topic modeling approach TET based on STMS approach [1], which can provide ranking of experts in different groups in an unsupervised way. It is generalized from a previous topic model ACT1 [15] from a single document "sub-group" (*no conferences influence*) to all publications of the conference "Group" (*conferences influence*). The intuition behind considering conferences as a group is explained with the help of an example in Fig. 1. A document denoted as a subgroup here, usually has a few semantically related words (as the total number of words in title is only "8") and authors (as the total number of authors is only "2") to a topic shown in Fig. 1, 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 the total number of words is as high as "439") and authors (as the total number of authors is as high as "95") to a topic as shown in Fig. 1. Subgroup is a subset of a group as highlighted in Fig. 1; 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 of a famous saying "A man is known by the company he keeps" and (2) accepted papers in highly ranked events are very carefully judged for relevance to the events areas of interests 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.

We empirically show that our proposed generalized time topic modeling approach (conference level (CL)) can clearly attain better results as compared to non-generalized time topic modeling approaches (document level (DL)) due to joint conferences influence and time information on the model performance. The solution produces promising and practical results.

Contributions from this work include: formalization of the temporal expert finding problem from CL with time information taken into account collectively, proposal of a generalized time topic modeling approach for the problem discussed, a method for unsupervised modeling for expert search requires only information about research papers and not other information such as impact factors of events where author has published, how many students they supervised or how many projects they have, and experimental verification of the effectiveness of the proposed approach using real world data. To the best of our knowledge, we are the first to deal with the temporal expert finding problem by proposing a generalized time topic modeling approach, which can capture word-to-word, word-to-author, author-to-author, word-to-event, author-to-event and event-to-event correlations by taking time factor into account, which is quite simple and effective.

The rest of the paper is organized as follows. In Section 2, we formalize the temporal expert finding problem. Section 3 provides related approaches and illustrates our proposed approach for temporal expert modeling with its details of parameters estimation and the derived model based on Bayes' Theorem. In Section 4, corpus, experimental settings, performance measures, baseline

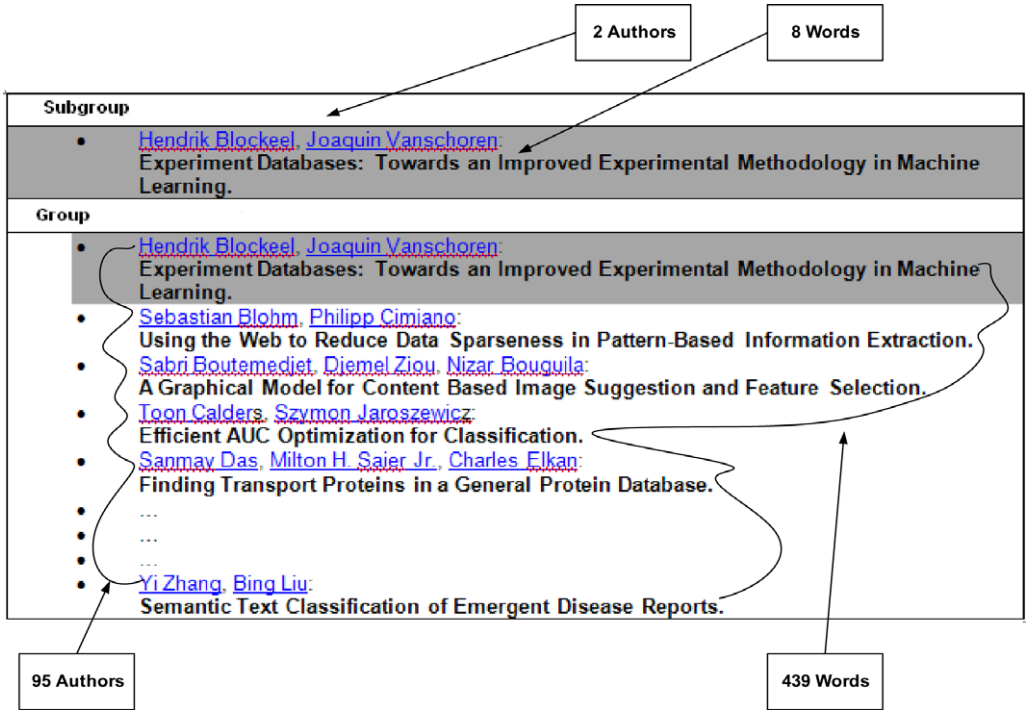

Fig. 1. An illustration of conferences influence (richer text semantics and relationships) with accepted papers by ECML/PKDD-2007.

Please cite this article in press as: A. Daud et al., Temporal expert finding through generalized time topic modeling, Knowl. Based Syst. (2010), doi:10.1016/j.knosys.2010.04.008

approaches with empirical studies and discussions about the results are given with practical applications of TET approach at the end of this section. Section 5 discusses related work and Section 6 brings this paper to a conclusion.

In the rest of the paper, we use the term event and conference interchangeably, conferences influence means "richer text semantics and relationships between authors" present between conferences. Additionally "super-document" or "group" means all the documents of one conference. In this work, non-generalized means document level (DL) and generalized means conference level (CL) approaches.

2. Temporal expert finding in the research community

Temporal expert finding addresses the task of discovering the people related to a specific knowledge domain for different time periods (e.g. years in this work). Expert finding became one of the biggest challenges in enterprises [12] and due to the dynamic spirits of writing and entry of new researchers in the research fields. 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 usually comprises of several words or token and a token is referred to as a collection of words as one term such as Data Mining.

We investigate the temporal expert finding problem by using a generalized time topic modeling approach. For example, each event accepts many papers every year written by different authors. To our interest, each publication contains some title words and names which usually cover most of the highly related sub-research areas of the conferences and authors, respectively. Events with their accepted papers on the basis of latent topics can help us to discover experts for different years. We think that latent topics based correlations between the authors publishing papers in specific events by considering time effects is an appropriate way for temporal expert discovery.

We denote an event c as a vector of N_c words based on the papers accepted by the event for a specific year y, an author r on the basis of his accepted paper(s), and formulize temporal expert finding problem as: Given an event c with N_c words having a stamp of year y, and $\mathbf{a}_c$ authors of an event c, discover most skilled persons of a specific domain for different years. Fig. 2 provides pictorial representation of the formulization of the temporal expert finding problem.

3. Temporal expert topic modeling

In this section, before describing our TET approach we briefly introduce the topic modeling idea followed by the basic topic modeling approach Latent Dirichlet Allocation (LDA) [9] and related approaches which are non-generalized Temporal-Author-Topic (TAT), non-generalized Author-Conference-Topic (ACT1) [15] and finally our generalized Temporal-Expert-Topic (TET) approach.

3.1. Topic modeling

Fundamental topic modeling assumes that there is a hidden topic layer $Z = \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \ldots, \mathbf{z}_t\}$ between the word tokens and the documents, where $\mathbf{z}_i$ denotes a latent topic and each document d is a vector of N_d words $\mathbf{w}_d$. A collection of D documents is defined by $D = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \ldots, \mathbf{w}_d\}$ and each word w_{id} is chosen from a vocabulary of size V. For each document, a topic mixture distribution is sampled and a latent topic Z is chosen with the probability of topic given document for each word with a word having generated probability of word given topic [9].

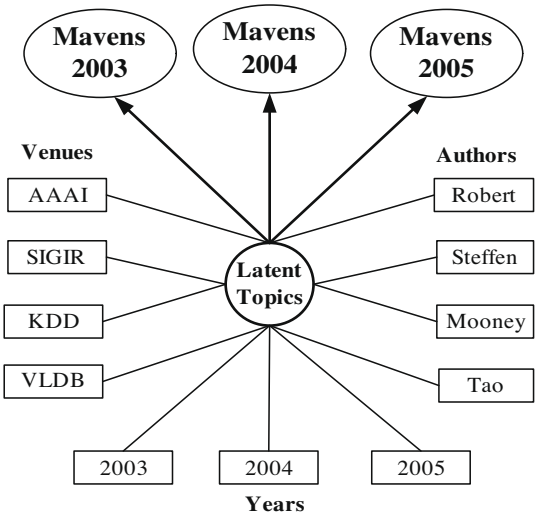

Fig. 2. Temporal expert finding.

3.2. Latent Dirichlet Allocation (LDA)

LDA [9] is the state-of-the-art topic model used for modeling documents by using a latent topic layer between them. It is a Bayesian network that generates a document using a mixture of topics. For each document d, a topic mixture multinomial distribution θ_d is sampled from Dirichlet α, and then a latent topic z is chosen and a word w is generated from topic-specific multinomial distribution Φ_z over words of a document for that topic. The following approaches are extensions of LDA

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

3.3. Non-generalized time topic modeling without conferences information

Previously, language model (LM) [8,32] and Probabilistic Latent Semantic Analysis (PLSA) [16,28] were proposed for expert discovery. These approaches ignored simultaneous modeling of time information and are only usually used for general purpose expert finding. In order to show the effectiveness of our approach a recently proposed non-generalized time topic modeling approach named as the Temporal-Author-Topic (TAT) [2] for finding dynamic authors interests is considered as one of the baselines. TAT simultaneously modeled time information of documents without considering conferences information. But one can say that conference information is very important when finding experts of specific area of research rather than just dynamic authors interests finding.

TAT [2] is a variation of ACT models [15], in which, each author from a set of K authors of a document is considered responsible for generating some topics of a document and in turn these topics generate the words and timestamps for that document. Formally, each author from a set of K authors of a document d 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 the document 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

Please cite this article in press as: A. Daud et al., Temporal expert finding through generalized time topic modeling, Knowl. Based Syst. (2010), doi:10.1016/j.knosys.2010.04.008

probability of the word w with year y for author r of document d is given as:

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

After we determined count matrices $W\times Z$ (word by topic), $Y\times Z$ (year by topic) and $Z\times R$ (topic by author) by using TAT, Eq. (9) is used for finding topically related experts for different years.

3.4. Non-generalized topic modeling with conferences information as token

Non-generalized time topic modeling approach TAT [2] used time information of documents without conferences information and also ignored conferences influence. A unified approach named Author-Conference-Topic (ACT1) was proposed by saying that authors and conferences are dependent on each other and should be modeled together [15]. ACT1 modeled conferences information of documents without considering conferences influence and simultaneous modeling of time information. But one can say that simultaneous modeling of time information is very important to acquire topics with almost similar topics over the years.

In ACT1 [15], each author from a set of K authors of a document is considered responsible for generating some latent topics of a document and in turn these topics generate the words and conferences stamps for that document. Formally, each author from a set of K authors of a document d 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 conference stamp for each word of the document 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 conference c for author r of document d is given as:

$$P(w,c|r,d,\emptyset,\Psi,\theta)=\sum_{z=1}^{T}P(w|z,\emptyset_z)P(c|z,\Psi_z)P(z|r,\theta_r) \quad (3)$$

3.5. Generalized time topic modeling with conferences influence

Non-generalized topic modeling approach ACT1 [15] 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. Proposed approach consists of two steps. In the first step, we use Semantics and Temporal Information based Maven Search (STMS) approach [1] for calculating the $W\times Z$ (word by topic), $Y\times Z$ (year by topic) and $Z\times R$ (topic by author) count matrices. In the second step, we derive a Bayes Theorem to determine topically related experts for different years.

STMS considers research papers as sub-entities of the events to model the influence of renowned and not-renowned events on the basis of authors' participation in similar events. Additionally, it considers time information to normalize the effect of words for different years and to get topic-year distribution. In this, an event is a composition of all documents words and the authors of its accepted publications with year as a stamp. Symbolically, for an event c (a super-document) one can write it as: $C=[\{(\mathbf{d}_1,\mathbf{a}_{d1})+(\mathbf{d}_2,\mathbf{a}_{d2})+(\mathbf{d}_3,\mathbf{a}_{d3})+\cdots+(\mathbf{d}_i,\mathbf{a}_{di})\}+y_c]$ where $\mathbf{d}_i$ is a word vector of a document published in an event, $\mathbf{a}_{di}$ is the author vector of $\mathbf{d}_i$ and y_c is the paper publishing year.

Non-generalized time topic modeling without conferences information, considers that an author is responsible for generating latent topics of the documents on the basis of semantics-based text information and authors correlations with time information. Non-generalized topic modeling with conferences information as token, considers that an author is responsible for generating latent topics of the documents on the basis of semantics-based text information and authors correlations. While, generalized time topic modeling with conferences influence, considers that an author is responsible for generating latent topics of the conferences on the basis of semantics-based text information and authors' correlations with consideration of time information (please see Fig. 3).

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:

$$P(w,y|r,c,\emptyset,\Psi,\theta)=\sum_{z=1}^{T}P(w|z,\emptyset_z)P(y|z,\Psi_z)P(z|r,\theta_r) \quad (4)$$

The generative process is as follows:

1. For each author $r=1,\ldots,K$ of event c.
   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 event c.Choose an author r uniformly from all authors $\mathbf{a}_c$.
   Choose a topic z from multinomial (θ_r) conditioned on r.
   Choose a word w from multinomial (Φ_z) conditioned on z.
   Choose a year y associated with word w from multinomial (Ψ_z) conditioned on z.

Gibbs sampling is used [6,29] for parameter estimation, which has two latent variables z and r; the conditional posterior distribution for z and r is given by:

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

where $z_i=j$ and $r_i=k$ represent the assignments of the ith word in an event to a topic j and author k respectively, $w_i=m$ represents the observation that the ith word is the mth word in the lexicon, $y_i=n$ represents ith year of paper publishing attached with the nth word in the lexicon and z_{-i} and r_{-i} represents all topic and author assignments not including the ith word. Furthermore, $n_{-i,j}^{(wi)}$ is the total number of words associated with topic j, excluding the current instance, $n_{-i,j}^{(yi)}$ is the total number of years associated with topic j, excluding the current instance and $n_{-i,j}^{(ri)}$ is the number of times author k is assigned to topic j, excluding the current instance, W is the size of the lexicon, Y is the number of years and R is the number of authors. "$\cdot$" Indicates summing over the column where it occurs and $n_{-i,j}^{(\cdot)}$ stands for number of all words and years that are assigned to topic z respectively, excluding the current instance.

During parameter estimation, the algorithm needs to keep track of $W\times Z$ (word by topic), $Y\times Z$ (year by topic) and $Z\times R$ (topic by author) count matrices. From these count matrices, topic-word

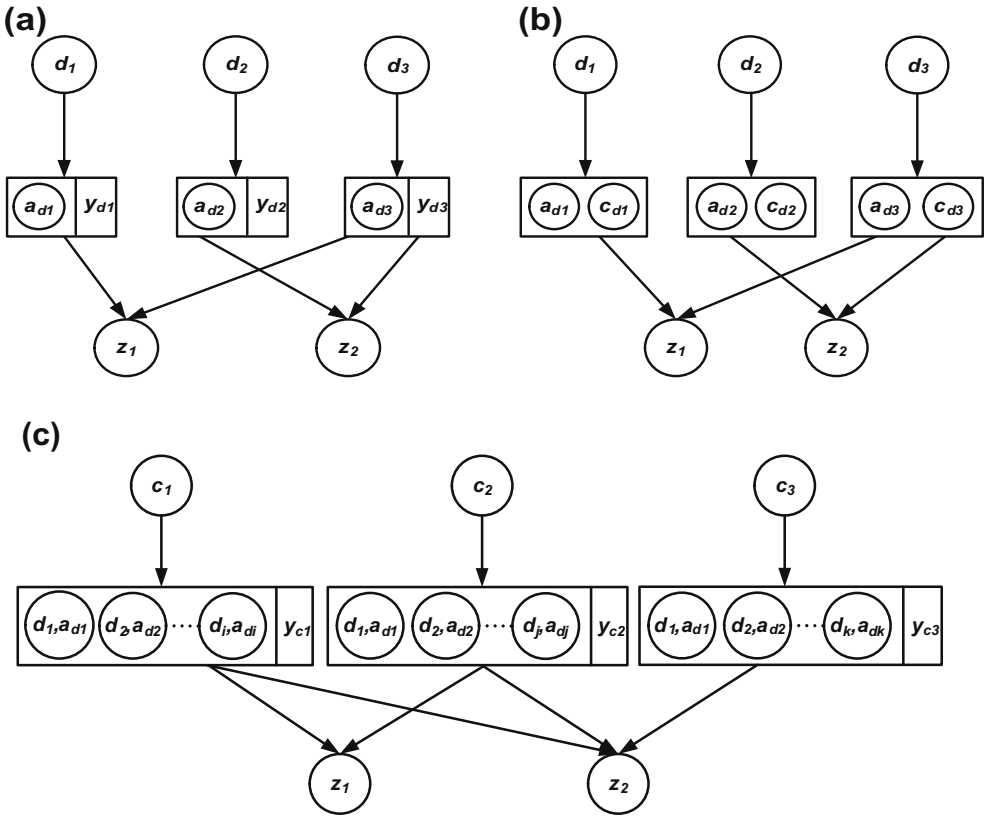

Fig. 3. Temporal expert topic modeling (a) TAT (non-generalized with time information and without conferences information), (b) ACT1 (non-generalized without time factor and with conferences information as token) and (c) TET (generalized with time factor and conferences influence) approaches.

distribution Φ, topic-year distribution Ψ and author-topic distribution θ can be calculated as:

$$\emptyset_{zw} = \frac{n^{(wi)}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + W\beta} \tag{6}$$

$$\Psi_{zy} = \frac{n^{(yi)}_{-i,j} + \gamma}{n^{(\cdot)}_{-i,j} + Y\gamma} \tag{7}$$

$$\theta_{rz} = \frac{n^{(ri)}_{-i,j} + \alpha}{n^{(ri)}_{-i,\cdot} + R\alpha} \tag{8}$$

where $\emptyset_{zw}$ is the probability of word w in topic z, Ψ_{zy} is the probability of year y for topic z and θ_{rz} is the probability of topic z for author r. These values correspond to the predictive distributions over new words w, new years' y and new topics z conditioned on w, y and z.

By deriving Bayes' Theorem, we can obtain the probability of an expert m given topic z and year y as:

$$P(m|z,y) = \frac{P(z,y|r)\cdot P(r)}{P(z,y)}, \quad \text{where}$$
$$P(z,y|r) = P(z|r)\cdot P(y|r) \quad \text{and} \quad P(y|r) = \sum_{z} P(y|z)\cdot P(z|r) \tag{9}$$

Here, for calculating $P(r)$ we simply used the number of publications of an author in a year. For more simplicity some works assume it uniform [19] and the propagation approach can be used to calculate it in a more complex way [17]. The effect similar to propagation effect is implicitly included in the proposed approach because of using Gibbs sampler, as more the papers the author has published more the probability he has of being ranked higher on the basis of attending event more times.

4. Experiments

4.1. Corpus

We downloaded five years paper corpus of conferences from DBLP database [11], by only considering conferences for which data was available for the years 2003–2007. In total, we extracted 112,317 authors, 90,124 papers, and combined them into a super document for 261 conferences per year. We then processed the corpus by (a) removing stop-words, punctuations and numbers, (b) converting to lower case the obtained words of papers, and (c) removing words and authors that appear less than three times in the corpus. This led to a vocabulary size of $V = 10{,}872$, a total of 572,592 words and 26,078 authors in the corpus. Fig. 4 shows the yearly data distribution for number of papers (D) and authors (R) in the conferences.

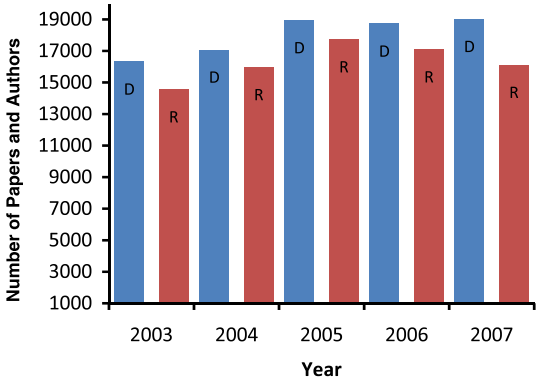

Fig. 4. Histogram illustrating data distribution.

Please cite this article in press as: A. Daud et al., Temporal expert finding through generalized time topic modeling, Knowl. Based Syst. (2010), doi:10.1016/j.knosys.2010.04.008

4.2. Parameter settings

In our experiments, for 100 topics Z the hyper parameters α, β and γ were set at 50/Z, 0.01, and 0.1. Topics are set at 100 by using perplexity [9] which is a standard measure for estimating the performance of probabilistic models with the lower being better, for the estimated topic models with human judgments of meaningful results.

4.3. Performance measures

Perplexity is usually used to measure the performance of latent topic based approaches which shows the generalization power of approach on the unseen data; however it cannot be a statistically significant measure when they are used for information retrieval [please see [21] for details]. We simply used entropy (square root of perplexity) and comprehensive analysis on the basis of real DBLP statistics for each author to evaluate the approach performance.

We used average entropy to measure the quality of discovered topics, which reveals the purity of topics, less intra-topic entropy is usually better (please see Eq. (10))

$$\text{Entrophy of (Topic)} = -\sum_{z} P(z)\log_2[P(z)] \tag{10}$$

To measure the performance in terms of precision and recall [21] is out of question due to unavailability of standard dataset and use of human judgments cannot provide unbiased answers for the performance evaluation. Consequently, we provide comprehensive (DBLP data Statistics) based comparison in Table 2. In it, we show how our proposed approach produced more precise results because of (1) top 10 experts in list published more in the World Level (World Class) conferences, (2) from top three conferences for each expert most of the time at least one of them is world level and (3) number of papers published by top 10 experts list for the topic are also greater.

4.4. Baseline approaches

We compared our proposed TET approach with two baseline approaches TAT [2] and ACT1 [15] and used the same number of topics for comparability. The numbers of Gibbs sampler iterations used for TAT and ACT1 are 1000 and parameter values are the same as the values used in [2,15], respectively. For TAT approach the process of finding the top 10 topically related experts for different years is same with TET. As ACT1 approach does not take time factor into account, so we divide dataset into five subsets by year, and run the model for each year to get experts for the different years.

4.5. Results and discussions

4.5.1. Experts for different years

We extracted and probabilistically ranked experts for different years related to a specific area of research on the basis of latent topics by using TET. The results are illustrated in Table 1. The

Table 1
An illustration of three discovered topics from 100 topics. Each topic is shown with the top 10 words (first column) and experts that have highest probability conditioned on that topic for each year (second to sixth column). Titles are our interpretation of the topics.

Words	Prob.	Year 2003		Year 2004		Year 2005		Year 2006		Year 2007	
		Experts	Prob.	Experts	Prob.	Experts	Prob.	Experts	Prob.	Experts	Prob.
Data Mining (DM) TET											
Topic 11											
Mining	0.163	2628	0.2	2628	0.2	2628	0.2	5135	0.2	2628	0.2
Data	0.100	5135	0.089	5135	0.149	5135	0.119	2628	0.148	5135	0.087
Clustering	0.043	5119	0.041	5119	0.044	5119	0.052	5119	0.043	4477	0.048
Frequent	0.028	4477	0.040	4477	0.035	2630	0.050	4477	0.036	118	0.027
Patterns	0.027	2630	0.036	2630	0.025	4477	0.041	2630	0.030	5119	0.016
Pattern	0.022	118	0.028	5014	0.024	118	0.024	10414	0.024	5014	0.014
Streams	0.019	4786	0.013	10307	0.020	10414	0.017	118	0.019	2630	0.013
Discovery	0.015	1659	0.013	118	0.020	5014	0.016	10307	0.013	5017	0.009
Preserving	0.015	5014	0.010	18258	0.016	5017	0.014	5014	0.013	1659	0.008
Ranking	0.012	5017	0.009	5017	0.016	4786	0.013	5002	0.012	10307	0.008
Bayesian Learning (XMLDB) TET											
Topic 81											
Learning	0.166	1844	0.2	1844	0.2	9636	0.2	1844	0.2	1844	0.2
Bayesian	0.045	9636	0.192	9636	0.144	1781	0.072	9636	0.192	9636	0.192
Markov	0.022	1781	0.087	11094	0.080	11094	0.071	3321	0.109	10279	0.080
Kernel	0.022	1637	0.074	8081	0.076	1844	0.069	21670	0.103	11025	0.076
Regression	0.018	11025	0.072	10827	0.076	8081	0.068	10279	0.100	3321	0.072
Reinforcement	0.018	3321	0.068	783	0.069	3321	0.063	1781	0.097	1781	0.055
Supervised	0.016	3289	0.059	1636	0.063	3289	0.057	18594	0.096	1636	0.051
Conditional	0.014	2247	0.053	2247	0.053	1636	0.051	783	0.087	16169	0.049
Inference	0.014	783	0.049	1781	0.052	1637	0.051	2247	0.086	11094	0.045
Random	0.013	9637	0.047	10279	0.050	20343	0.046	11094	0.082	21670	0.044
XML Databases (XMLDB) TET											
Topic 74											
Data	0.109	350	0.2	5258	0.2	9382	0.2	5258	0.2	5258	0.2
Xml	0.064	5258	0.164	9382	0.143	5258	0.162	9382	0.123	350	0.192
Query	0.048	4808	0.113	4808	0.122	350	0.160	350	0.078	9382	0.080
Databases	0.038	5291	0.085	350	0.121	5291	0.112	4808	0.065	4808	0.076
Database	0.036	9382	0.075	5291	0.092	4808	0.093	16464	0.059	9381	0.072
Queries	0.03	9963	0.060	4457	0.060	2621	0.090	9963	0.056	4775	0.055
Processing	0.026	4455	0.054	4870	0.051	18023	0.070	2621	0.054	14090	0.051
Relational	0.021	2621	0.052	4775	0.050	18396	0.064	18023	0.049	18396	0.049
Indexing	0.016	4775	0.050	5320	0.046	4457	0.062	2316	0.040	4820	0.045
Integration	0.015	2005	0.046	4490	0.042	16464	0.056	5101	0.039	9963	0.044

Please cite this article in press as: A. Daud et al., Temporal expert finding through generalized time topic modeling, Knowl. Based Syst. (2010), doi:10.1016/j.knosys.2010.04.008

words associated with each topic provide a meaningful description of a specific area of research. For example, topic # 74 "XML Databases (XMLDB)" shows quite specific words when a person is finding databases experts specifically for XML databases. Other topics Data Mining and Bayesian Learning shown in Table 1 are also quite descriptive and precise. The experts associated with each topic are quite representative, e.g. we analyzed the data and found that in case of topic 11 "Data Mining" and for others topics top ranked experts are well known in their respective fields and published usually in high class conferences, specifically from top three conferences for each expert at least one is world class. Here it is necessary to mention that these rankings are just based on the five years (2003–2007) DBLP data of only related conferences to just show conferences influence for TET in comparison with TAT and ACT1 and ids are mentioned instead of names for anonymity.

Gibbs sampling algorithm requires significant processing time which is not practice for large datasets. We usually need to quickly find the topics and experts for new events that are not contained in the training corpus. For this purpose, we can apply Eq. (5) only on the word tokens and authors of the new event each time temporarily updating the count matrices of (word by topic), (year by topic) and (topic by author). The resulting assignments of words to topics can be saved after a few iterations (10 in our simulations) and then Eq. (9) is used to obtain experts for different years.

Table 2a
Temporal expert finding comparison between our proposed and baseline approaches for DM topic related top 10 experts. Here the top 10 experts related to a topic for a year with the top three conferences shown in which they published and number of papers they have published in that year.

Experts	Top three conferences	TP	Experts	Top three conferences	TP	Experts	Top three conferences	TP
2003 Data Mining (TET)			*2003 Data Mining (ACT1)*			*2003 Data Mining (TAT)*		
2628	**WL (ICDE)**, NL (ISCAS, ICDM)	33	4477	**WL (ICDE, KDD, SIGMOD)**	19	4477	**WL (ICDE, KDD, SIGMOD)**	19
5135	NL (BIBE, DAWAK, GRC)	9	2681	NL (ICDM, IDEAS, MDM/KDD)	10	4921	NL (SEBD, PKDD, ICDM)	4
5119	**WL (ICDE, SIGMOD)**, NL (DASFAA)	13	5018	NL (ICDM, ADMA, APIN)	6	2067	NL (SIGPlan, CompSAC, CW)	20
4477	**WL (ICDE, KDD, SIGMOD)**	19	2231	NL (AAI, ADC, AI)	5	25858	**WL (KDD)**, NL (PAKDD, SDM)	4
2630	**WL (KDD)**, NL (ICDM, IPDPS)	11	1660	**WL (ICDE, SIGMOD)**, NL (ICDM)	12	2587	NL (ICDM, CIKM, SDM)	7
118	**WL (SIGIR)**, NL (ICDM, CIKM)	12	2630	**WL (KDD)**, NL (ICDM, IPDPS)	11	212	**WL (KDD,SIGMOD, VLDB)**	13
4786	**WL (KDD)**, NL (ICDM, SDM)	14	8642	NL (ICEIS, ICWI, IKE)	9	2630	**WL (KDD)**, NL (ICDM, IPDPS)	11
1659	**WL (KDD)**, NL (ICDM, SDM)	6	323	NL (CIKM, PAKDD, ICDCS)	19	4399	NL (ISAAC, DS, WG)	7
5014	**WL (SIGIR, WWW)**, NL (ICDM)	8	5325	**WL (KDD)**, NL (ICTAI, DASFAA)	9	2239	**WL (ICDE)**, NL (DASFAA, WAIM)	23
5017	NL (ICEIS, SAC, IRI)	14	8737	NL (CIKM, PAKDD, IDEAS)	7	8737	NL (CIKM, PAKDD, IDEAS)	7
2004 Data Mining (TET)			*2004 Data Mining (ACT1)*			*2004 Data Mining (TAT)*		
2628	**WL (ICDE)**, NL (ISCAS, ICDM)	58	5017	NL (ICEIS, SAC, IRI)	27	4477	**WL (ICDE, KDD, SIGMOD)**	26
5135	NL (BIBE, DAWAK, GRC)	10	1661	**WL (ICDE, KDD)**, NL (ICDM)	36	25858	**WL (KDD)**, NL (PAKDD, SDM)	11
5119	**WL (ICDE, SIGMOD)**, NL (DASFAA)	26	9175	NL (ICCSA, AINA, iiWAS)	40	2587	NL (ICDM, CIKM, SDM)	13
4477	**WL (ICDE, KDD, SIGMOD)**	26	1832	**WL (KDD)**, NL (ICTAI, ICDM)	17	5017	NL (ICEIS, SAC, IRI)	27
2630	**WL (KDD)**, NL (ICDM, IPDPS)	11	4490	NL (DEXA, DASFAA, CIKM)	14	4921	NL (SEBD, PKDD, ICDM)	5
5014	**WL (SIGIR, WWW)**, NL (ICDM)	22	2239	**WL (ICDE)**, NL (DASFAA, WAIM)	13	2630	**WL (KDD)**, CL (ICDM, IPDPS)	11
10307	**WL (KDD)**, NL (ICDM, SDM)	7	4857	**WL (ICDE, VLDB, PODS)**	7	9382	NL (WAIM, APWEB, FSKD)	19
118	**WL (SIGIR)**, NL (ICDM, CIKM)	10	4477	**WL (ICDE, KDD, SIGMOD)**	26	2067	NL (SIGPlan, CompSAC, CW)	28
18258	NL (ICPR, ICAPR, IDEAL)	12	2630	**WL (KDD)**, NL (ICDM, IPDPS)	11	10307	**WL (KDD)**, NL (ICDM, SDM)	7
5017	NL (ICEIS, SAC, IRI)	27	8077	NL (IWANN, ICANN, IJON)	13	212	**WL (KDD,SIGMOD, VLDB)**	14
2005 Data Mining (TET)			*2005 Data Mining (ACT1)*			*2005 Data Mining (TAT)*		
2628	**WL (ICDE)**, NL (ISCAS, ICDM)	61	1661	**WL (ICDE, KDD)**, NL (ICDM)	50	4477	**WL (ICDE, KDD, SIGMOD)**	30
5135	NL (BIBE, DAWAK, GRC)	22	4477	**WL (ICDE, KDD, SIGMOD)**	30	4921	NL (SEBD, PKDD, ICDM)	7
5119	**WL (ICDE, SIGMOD)**, NL (DASFAA)	23	212	**WL (KDD, ICDE, SIGMOD)**	16	2630	**WL (KDD)**, NL (ICDM, IPDPS)	17
2630	**WL (KDD)**, NL (ICDM, IPDPS)	17	965	**WL (VLDB)**, NL (SODA, CoRR)	26	2067	NL (SIGPlan, CompSAC, CW)	35
4477	**WL (ICDE, KDD, SIGMOD)**	30	24720	**WL (KDD)**, NL (BIOKDD, ICDM)	18	212	**WL (KDD,SIGMOD, VLDB)**	16
118	**WL (SIGIR)**, NL (ICDM, CIKM)	23	2292	NL (AJAI, PAKDD, AAI)	8	5017	NL (ICEIS, SAC, IRI)	31
10414	**WL (ICDE)**, NL (PAKDD, ICDM)	3	1660	**WL (ICDE, SIGMOD)**, NL (ICDM)	15	9382	NL (WAIM, APWEB, FSKD)	17
5014	**WL (SIGIR, WWW)**, NL (ICDM)	20	7771	NL (HIS, ICCS, COR)	13	2587	NL (ICDM, CIKM, SDM)	13
5017	NL (ICEIS, SAC, IRI)	31	4982	NL (PAKDD, ADC, CAiSE)	11	25858	**WL (KDD)**, NL (PAKDD, SDM)	3
4786	**WL (KDD)**, NL (ICDM, SDM)	18	4490	NL (DEXA, DASFAA, CIKM)	14	4786	WL (KDD), NL (ICDM, SDM)	18
2006 Data Mining (TET)			*2006 Data Mining (ACT1)*			*2006 Data Mining (TAT)*		
5135	NL (BIBE, DAWAK, GRC)	23	4477	**WL (ICDE, KDD, SIGMOD)**	34	4477	**WL (ICDE, KDD, SIGMOD)**	34
2628	**WL (ICDE)**, NL (ISCAS, ICDM)	91	1661	**WL (ICDE, KDD)**, NL (ICDM)	43	25858	**WL (KDD)**, NL (PAKDD, SDM)	16
5119	**WL (ICDE, SIGMOD)**, NL (DASFAA)	32	2630	**WL (KDD)**, NL (ICDM, IPDPS)	21	4921	NL (SEBD, PKDD, ICDM)	8
4477	**WL (ICDE, KDD, SIGMOD)**	34	2644	**WL (KDD)**, NL (ICDM, SDM)	19	2630	**WL (KDD)**, NL (ICDM, IPDPS)	21
2630	**WL (KDD)**, NL (ICDM, IPDPS)	21	323	NL (CIKM, PAKDD, ICDCS)	23	5017	NL (ICEIS, SAC, IRI)	30
10414	**WL (ICDE)**, NL (PAKDD, ICDM)	10	2621	NL (DEXA, ICWL, MDM)	16	4968	NL (ICTAI, SDM, WebKDD)	12
118	**WL (SIGIR)**, NL (ICDM, CIKM)	41	5135	NL (BIBE, DAWAK, GRC)	23	212	**WL (KDD,SIGMOD, VLDB)**	20
10307	**WL (KDD)**, NL (ICDM, SDM)	13	10282	**WL (KDD, ICDE)**, NL (SDM)	11	10326	**WL (KDD)**, NL (SDM, PKDD)	9
5014	**WL (SIGIR, WWW)**, NL (ICDM)	21	2628	**WL (ICDE)**, NL (ISCAS, ICDM)	91	9382	NL (WAIM, APWEB, FSKD)	20
5002	**WL (KDD)**, NL (SDM, ICDM)	12	2247	NL (ADC, ICDT, WISE)	12	10307	**WL (KDD)**, NL (ICDM, SDM)	13
2003 Data Mining (TET)			*2003 Data Mining (ACT1)*			*2003 Data Mining (TAT)*		
2628	**WL (ICDE)**, NL (ISCAS, ICDM)	92	4477	**WL (ICDE, KDD, SIGMOD)**	45	4477	**WL (ICDE, KDD, SIGMOD)**	45
5135	NL (BIBE, DAWAK, GRC)	22	1832	**WL (KDD)**, NL (ICTAI, ICDM)	24	25858	**WL (KDD)**, NL (PAKDD, SDM)	12
4477	**WL (ICDE, KDD, SIGMOD)**	45	9377	NL (SBBD, BDA, DAWAK)	6	212	**WL (KDD,SIGMOD, VLDB)**	24
118	**WL (SIGIR)**, NL (ICDM, CIKM)	42	4793	**WL (KDD, VLDB, SIGMOD)**	24	25817	NL (RAID, ACSAC, NADD)	11
5119	**WL (ICDE, SIGMOD)**, NL (DASFAA)	14	2628	**WL (ICDE)**, NL (ISCAS, ICDM)	92	5017	NL (ICEIS, SAC, IRI)	28
5014	**WL (SIGIR, WWW)**, NL (ICDM)	18	24720	**WL (KDD)**, NL (BIOKDD, ICDM)	7	10326	**WL (KDD)**, NL (SDM, PKDD)	11
2630	**WL (KDD)**, NL (ICDM, IPDPS)	10	9894	NL (DG.O, ICS, DL)	3	4921	NL (SEBD, PKDD, ICDM)	10
5017	NL (ICEIS, SAC, IRI)	28	1996	**WL (PODS)**, NL (APBC,DPBL)	8	4968	NL (ICTAI, SDM, WebKDD)	6
1659	**WL (KDD)**, NL (ICDM, SDM)	12	8737	NL (CIKM, PAKDD, IDEAS)	8	8737	NL (CIKM, PAKDD, IDEAS)	8
10307	**WL (KDD)**, NL (ICDM, SDM)	6	25261	NL (WAIM, RSKT, CSDA)	2	4399	NL (ISAAC, DS, WG)	13

Please cite this article in press as: A. Daud et al., Temporal expert finding through generalized time topic modeling, Knowl. Based Syst. (2010), doi:10.1016/j.knosys.2010.04.008

4.5.2. DBLP data statistics based comparison

To show the dominancy of our proposed approach over the baseline approaches, we provide comparison of all years for DM topic by using DBLP database [11] 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 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 topid**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 [11] and categorized them as WL (bold font in Top 3 Conferences column) and NL (normal font in Top 3 Conferences column) in Table 2a. Total Papers (TP) column shows number of papers published in a given year by the expert in all conferences.

One can see in Table 2b a summary of Table 2a; firstly, for year 2003 of TET from top 10 experts 12 times papers are published in WL conferences with total number of 139 papers, for year 2003 of ACT1 from top 10 experts 7 times papers are published in WL conferences with total number of 107 papers, and for year 2003 of TAT from top 10 experts 9 times papers are published in WL conferences with total number of 115 papers. Twelve times WL for TET is greater than 7 times WL for ACT1 and 9 times of WL for TAT.

Secondly, eight experts from top 10 shown for TET at least have OneWL conference related to an expert in top three conferences, four experts from top 10 for ACT1 at least have OneWL conference related to an expert in top three conferences, and five experts from top 10 for TAT at least have OneWL conference related to an expert in top three conferences. Eight experts OneWL for TET are greater than four experts for ACT1 and five experts for TAT. Thirdly, 139 TP for TET are greater than 107 TP for ACT1 and 115 TP for TAT. It clearly shows that experts found by TET approach are better, as they published more in WL conferences, more experts in the top 10 lists have published at least in one OneWL and experts published more papers as compared to ACT1 and TAT. The above situation is also true for years 2004, 2005, 2006 and 2007.

Thirdly, Table 2b shows that the average number of times experts publishing in WL 12 for TET is greater than WL 10 of ACT1 and WL 8.8 of TAT, 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 and average OneWL 4.6 of TAT, which supports our hypothesis that our approach can discover more precise experts who published more in WL conferences than experts discovered by other approaches.

One can say that if someone is an expert in some area of research he should have at least one world class conference among his/her top three publishing conferences. Additionally, the average number of papers for TET approach for the top 10 experts is 236.6 which is greater than the average number of papers for ACT1 approach 204.8 and TAT approach 162.8, which shows the proposed approach acquiring more accurate results.

The results presented in Table 2b show that TET approach outperformed ACT1 and TAT approaches due to its ability to simultaneously capture conferences influence with time information, and TAT approach performed poorer than ACT1 because of not considering conferences information even as token.

4.5.3. Entropy based comparison

Fig. 5 provides a comparison between TET, TAT and ACT1 approaches in terms of entropy. It shows the average entropy curve of topic-word distribution for all topics calculated by using Eq. (10). Lower entropy for different number of topics $T = 100$, 150, 200 shows that more dense topics are produced by the approach. Here, TAT has a little less entropy than TET and much less than ACT1, while TAT and TET both have less entropy than ACT1. The experimental results in Table 2 clearly show that the TET approach is better than the TAT and ACT1 approaches even TAT has a little less entropy (means dense topics) than TET. It matches with the description given in [21] that perplexity or entropy is not statistically significant evaluation measures when topic models are used for object ranking in information retrieval. As in this case TAT produced more dense topics (low entropy) than that of TET and ACT1 but performed poorer than both for temporal expert finding.

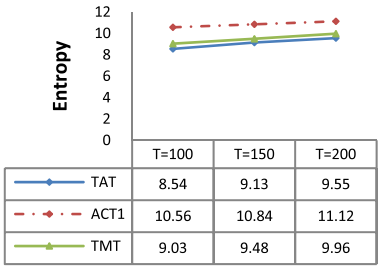

Fig. 5. Average entropy curves as a function of different number of topics.

Additionally, the average entropy curve for different number of topics for all approaches is pretty much smooth, which indicates that these approaches are not sensitive to the number of topics.

4.5.4. Exchangeability of topics

ACT1 model [15] does not simultaneously consider time information (independently modeled topics for each year) with the text and authors' information, which results in exchangeability of topics problem. It means that there is no fixed order of topics for different runs of the algorithm. For example, a topic z_i in the first run of the algorithm cannot be considered same as topic z_i in the different runs of the Gibbs sampling algorithm [3]. Consequently, when we run ACT1 for finding dynamic research interests for five years individually that resulted into two main problems. Firstly, the topics numbers were not similar for different years. Which may need one to find some method to first map similar topics for different years and obtain further results for knowledge based system. Secondly, the probabilistically related words are not exactly leading to same area of interest for different years. These problems result in having topically related biased researchers for the topics, as an example the Bayesian learning topic is discussed here. Table 3 provides the topic words for different years obtained by ACT1 for "Bayesian" learning: topic. It enlightens the problem of not having similar topic number and probabilistic words for each year because of modeling independently for each year. For example, word "machine" is important for Bayesian learning topic but is missing for year 2004, word "Bayesian" is important but missing for years 2003 and 2007, word "semi" is present for only years 2004, 2005 and 2007, and word "hybrid" is only present for the year 2006. Non-presence of important words like machine, Bayesian and presence of unimportant words like hybrid in Bayesian learning topic results in finding experts misleads to the similar area of interest for different years.

4.6. Applications of TET approach

4.6.1. Temporal social network of Zoubin Ghahramani

TET approach can also be used for dynamic correlation discovery between experts for different years, as compared to only discovering static authors' correlations [23]. To illustrate how it can be used in this respect, distance between experts i and j is

Please cite this article in press as: A. Daud et al., Temporal expert finding through generalized time topic modeling, Knowl. Based Syst. (2010), doi:10.1016/j.knosys.2010.04.008

Table 2b
Summary of Table 2a, here WL means World Class conference, OneWL means at least one conference is WL in top three conferences related to an expert and TP means total number of papers for the top 10 topically related authors to a topic.

Year	WL (TAT)	WL (ACT1)	WL (TET)
2003	9	7	12
2004	9	11	11
2005	8	12	12
2006	10	10	13
2007	8	10	12
Average	8.8	10	12
	OneWL (TAT)	OneWL (ACT1)	OneWL (TET)
2003	5	4	8
2004	5	6	7
2005	4	6	8
2006	6	6	9
2007	3	6	8
Average	4.6	5.6	8
	TP (TAT)	TP (ACT1)	TP (TET)
2003	115	107	139
2004	161	204	209
2005	187	201	248
2006	183	293	298
2007	168	219	289
Average	162.8	204.8	236.6

calculated by using Eq. (11) for author-topic distribution of different years

$$sKL(i,j) = \sum_{z=1}^{T} \left[\theta_{iz} \log \frac{\theta_{iz}}{\theta_{jz}} + \theta_{jz} \log \frac{\theta_{jz}}{\theta_{iz}} \right] \quad (11)$$

We calculated the dissimilarity between the authors; smaller dissimilarity value means higher correlation between the experts. Table 4 shows topically related people with Zoubin Ghahramani for different years. We selected Zoubin Ghahramani as he is famous person and has mature social network instead in comparison to a new researcher who may not have mature social network based on his research interests. Here, it is obligatory to mention that the top 10 people related to Zoubin Ghahramani are not necessarily the experts who have co-authored with him mostly, but rather are the people who tend to produce most words for the same topics with him for a specific year and usually attended the same events. Again the results are quite promising and realistic as most of the people related to Zoubin Ghahramani for different years are also related to BL topic (e.g. Andrew Y. Ng, James T. Kwok and Raymong J. Mooney) in Table 1 or are well known in this area of research.

4.6.2. Temporal interests

Now by using TET we will show topic-wise experts temporal interests. In Fig. 6a on the left side, for DM topic Hans-Peter Kriegel's topic interest curve rises a bit from 2003 to 2004 and is very much smoother from year 2004 to 2006 but suddenly falls for the year 2007. This situation matches well with the DBLP data statistics as he published papers are 13, 26, 23, 32, 14 in order of years (2003–2007) and in 2007 he published only 14 papers, which are 18 papers less than year 2006, as a result his topic interest curve decreased rapidly. For a similar topic Wei Fan has an almost stable topic interest curve which has no prominent rising or falling trends. This situation matches well with his published papers 6, 9, 6, 10, 12 in order of years (2003–2007) in DBLP data statistics. For DM topic Jiawei Han has a slowly increasing topic interest curve from year 2003–2007. This situation matches well with the DBLP data statistics as his published papers are 19, 26, 30, 34, 45 in order of years (2003–2007) with the number of publications per year are increasing steadily.

In Fig. 6b on the right side, for XMLDB topic Kian-Lee Tan's publishing rate decreased from year 2003 to 2007 which is shown by the trend line. This situation matches well with the DBLP data statistics as his published papers are 29, 27, 25, 21, 14 in order of

Table 3
Exchangeability of topics problem.

Topic "90" year 2003		Topic "59" year 2004		Topic "3" year 2005		Topic "11" year 2006		Topic "83" year 2007	
Learning	0.182262	Learning	0.221903	Learning	0.226613	Learning	0.208064	Learning	0.222570
Machine	0.033243	**Semi**	0.020124	**Bayesian**	0.033181	**Machine**	0.028394	Kernel	0.024824
Reinforcement	0.012730	**Bayesian**	0.018496	Supervised	0.016615	Active	0.017700	Active	0.022954
Programming	0.006697	Markov	0.016869	Reinforcement	0.015153	**Bayesian**	0.014491	Classification	0.022019
Function	0.006697	Reinforcement	0.016327	**Semi**	0.013691	Reinforcement	0.011818	**Machine**	0.020616
Environment	0.006697	Supervised	0.013072	Bayes	0.008819	**Hybrid**	0.006470	Supervised	0.020616
Metrics	0.006094	Function	0.009818	Active	0.008332	Automated	0.005401	**Semi**	0.018279
Discovery	0.005490	Classifiers	0.008733	**Machine**	0.007357	Concurrent	0.004866	Reinforcement	0.015006
Robust	0.005490	Models	0.007648	Ranking	0.006383	Developing	0.004866	Unsupervised	0.006592
Structure	0.004887	Network	0.006563	Intelligent	0.005896	Simulation	0.004866	Statistical	0.006124

Table 4
Top 10 associated authors with Zoubin Ghahramani for different years.

2003		2004		2005		2006		2007	
Zoubin Ghahramani									
Michael I. Jordan	0.138	Andrew Y. Ng	0.253	Manfred K. Warmuth	0.556	Andrew Y. Ng	0.516	Michael I. Jordan	0.249
Andrew Y. Ng	0.174	Manfred K. Warmuth	0.280	James T. Kwok	0.609	Raymond J. Mooney	0.531	Rich Caruana	0.261
Yoram Singer	0.180	Zhi-Hua Zhou	0.290	Andrew Y. Ng	0.610	Rich Caruana	0.587	Sanjay Jain	0.283
Shie Mannor	0.180	Raymond J. Mooney	0.291	Raymond J. Mooney	0.695	Pieter Abbeel	0.676	Michael L. Littman	0.285
Zhi-Hua Zhou	0.188	John Langford	0.371	Zhi-Hua Zhou	0.917	Qiang Yang	0.761	Shie Mannor	0.303
Max Welling	0.201	Michael I. Jordan	0.385	Pieter Abbeel	0.926	Satinder P. Singh	0.785	Dan Roth	0.311
Raymond J. Mooney	0.208	Pieter Abbeel	0.415	John Langford	1.033	Zhi-Hua Zhou	0.791	Sridhar Mahadevan	0.312
Marcus Hutter	0.221	Qiang Yang	0.500	Volker Tresp	1.154	Harry Zhang	0.802	Tony Jebara	0.318
Sanjay Jain	0.222	Pedro Domingos	0.504	Sridhar Mahadevan	1.202	Michael L. Littman	0.837	Yoram Singer	0.321
Pieter Abbeel	0.2539	Tong Zhang	0.506	Robert E. Schapire	1.255	Max Welling	0.847	Max Welling	0.323

Please cite this article in press as: A. Daud et al., Temporal expert finding through generalized time topic modeling, Knowl. Based Syst. (2010), doi:10.1016/j.knosys.2010.04.008

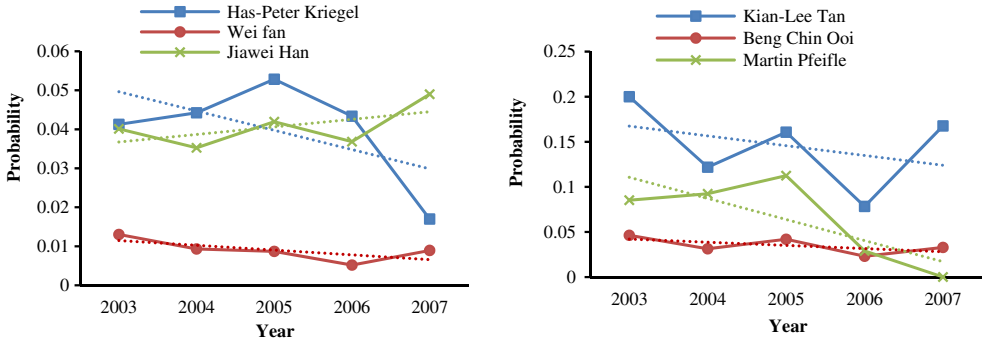

Fig. 6. Topic-wise temporal experts interests for (a) Data Mining (left) and (b) Bayesian learning (right).

years (2003–2007) are decreased especially for 2006 and 2007 21 and 14 papers, due to which the trend line is dropping slowly. For a similar topic Beng Chin Ooi has stable topic interest as topic interest curve has no prominent rising or falling trend from year 2003 to 2007. This situation matches well with his published papers 21, 17, 21, 17, 11 in order of years (2003–2007) in DBLP data statistics. Here one can say that for year 2007 number of papers are 11 which are less than 17 (number of papers in 2006) even then trend line is quite smooth. It is necessary to mention that here some time even a person published less number of papers but he can still produce more semantically related words, which nullifies the impact of a lesser number of publications by showing his more focused interest for that topic. For similar topic Martin Pfeifle topic interest curve is continuously decreasing from 2003 to 2005, then it falls suddenly for year 2006 and eventually it is almost touching zero for year 2007. This situation matches well with the DBLP data statistics as his published papers are 8, 14, 12, 5, 1 in order of years (2003–2007) with his number of publications per year decreasing quickly for the years 2006 and 2007. Temporal interests results explained above clearly show the effectiveness of our proposed approach and are well justified.

5. Related work

5.1. Temporal expert finding

Text Retrieval Conference (TREC) has provided a common platform for researchers to empirically assess approaches for expert finding. Different approaches have been proposed to handle this problem. In particular, Cao et al. [32] proposed a two-stage language model which combines a co-occurrence model to retrieve documents related to a given query, and a relevance model for expert search in those documents. Balog et al. [19] proposed a model which models candidates using its support documents by using language modeling framework. Later, he proposed several advance models to study expert finding problem in sparse data environments [20]. Different models for expert finding are compared by Petkova and Croft [14], which were probabilistically equivalent but their difference lie in the independent assumptions of models. Liu et al. [30] studied a weighted directed co-author network and proposed AuthorRank algorithm for ranking authors. Nie et al. [34] proposed PopRank link analysis model a variation of language model for object ranking within a specific domain. Expertise search is performed in time varying social networks by proposing a Temporal Random Walk algorithm [33]. Zhang et al [16] discussed preceding approaches limitation of not capturing semantics-based information and proposed a mixture model (MM) based on the Probabilistic Latent Semantic Analysis (PLSA). MM used a latent topic layer between authors' documents and the query. Expertise topic modeling for matching reviewers with the papers is investigated without considering conferences information [13]. Recently, dependency between conferences and authors is argued and a unified Author-Conference-Topic (ACT1) model is proposed to capture the combined influence of conferences and authors for expertise search [15]. To the limitation of their work conferences information is only used as a token which became the reason for ignoring conferences influence.

Time information usage has become important because of highly dynamic Web for most of the knowledge discovery tasks. Especially, Alonso et al. [24] argued that most of the existing information retrieval systems do not consider time information and have shown some areas in which one can benefit by exploiting time information. Most of the existing approaches ignored semantics-based information or were focused on finding general experts by using semantics-based information. In semantics-based topic modeling approaches time was ignored or modeling was done from document level which became the reason for ignoring conferences influence. Proposed TET approach incorporates both conferences influence and time information for temporal expert finding, simultaneously.

5.2. Topic modeling

Automatic topic extraction from data has become interesting and active area of research and topic modeling is mostly used for fulfilling this task. A few efforts have been made for topic extraction by using hard clustering methods to cluster documents into one specific group based on similar semantic contents [4,5]. Practically a document can have more than one topic e.g. this paper at least has two topics; which are temporal expert finding and topic modeling. Therefore Probabilistic Latent Semantic Analysis (PLSA) [28] was proposed as a probabilistic alternative to provide soft clusters of documents on the basis of which one document can belong to two or more groups. PLSA was generative only at words level but not at the documents level, which is considered its main limitation, as it has no simple way to make a prediction for new document added to corpus. Consequently, a probabilistic topic model Latent Dirichlet Allocation (LDA) was proposed [9], which was generative at both words and documents level by using Dirichlet prior. It enables LDA to predict topics for new document added to the corpus. Later, LDA was extended to Author-Topic model [23] for modeling authors' interests on the basis of latent topics without conferences information.

Jianping and Shiyong [18] discussed that the number of words in a document is greatly different from that in other documents and it is difficult to perform topic transition analysis based on current topic models. Consequently, a variable space hidden Markov model (VSHMM) is proposed to represent the topics, and several operations based on space computation are presented. The task of selecting relevant features for unsupervised clustering is

Please cite this article in press as: A. Daud et al., Temporal expert finding through generalized time topic modeling, Knowl. Based Syst. (2010), doi:10.1016/j.knosys.2010.04.008

considered difficult due to the absence of class labels that can be useful for search. Therefore, a new mixture model method is proposed for unsupervised soft text clustering, named multinomial mixture model with feature selection (M3FS) [22].

Time topic modeling has become important because of highly dynamic Web. Blei and Lafferty [10] proposed dynamic topic model (DTM) which can capture the evolution of topics in sequentially organized data. Later, Wang and McCallum proposed Topic over Time (TOT) [31] which used beta distribution to draw time stamps of years to investigate the evolution of topics. Nallapati et al. [26] proposed a Multiscale Topic Tomography model (MTTM) to model the evolution of topics over time. DTM, TOT and MTTM all of them can be considered as time topic models but they did not consider authors and conferences influence, which motivated us to propose TET.

6. Conclusions

This study deals with the problem of temporal expert finding by simultaneously modeling conferences influence and time information. The effect of generalizing topic modeling approach from document level to conference level is investigated. We conclude that it is significant to use conferences influence and time information together as our generalized time topic modeling approach discovered experts for different years are more precise than non-generalized topic modeling approaches. TET approach is used to effectively find temporal expert correlations and topic-wise interests; clear and useful results are obtained. We also show that entropy (or perplexity (square of entropy)) cannot be a significant measure for performance evaluation when topic model is used for specific problem such as ranking for information retrieval. As in this work, the TAT approach has lesser entropy than both TET and ACT1 approaches; even then it performed poorer than both because it has no conferences information which is important for temporal expert finding problem.

We also discussed the exchangeability of topics problem with ACT1 by not simultaneously modeling time information and conclude that simultaneous modeling of time information is important for temporal expert finding. Empirical results and detailed analysis show that TET outperformed TAT and ACT1 approaches. In general our proposed approach can be applied to blogs dataset for finding influential bloggers and news dataset for finding expert news reporters.

Acknowledgements

The work is supported by the National Natural Science Foundation of China under Grant (60973102, 60703059), Chinese National Key Foundation Research and Development Plan under Grant (2007CB310803) and Higher Education Commission, Islamabad, Pakistan for providing scholarship to the first author the main contributor of this work. We are thankful to Jie Tang for sharing his topic modeling codes.

References

[1] A. Daud, J. Li, L. Zhou, F. Muhammad, A generalized topic modeling approach for Maven search, in: Proceedings of International Asia-Pacific Web Conference and Web-Age Information Management (APWeb-WAIM), Suzhou, China, 2–4 April, 2009.

[2] A. Daud, J. Li, L. Zhou, F. Muhammad, Exploiting temporal authors interests via temporal-author-topic modeling, in: Proceedings of International Conference on Advanced Data Mining and Applications (ADMA), Beijing, China, 17–19 August, 2009.

[3] A. Daud, J. Li, L. Zhou, F. Muhammad, Knowledge discovery through parametric directed probabilistic topic models – A survey, Journal of Frontiers of Computer Science in China (FCS), (2010), in press, doi:10.1007/s11704-009-0062-y.

[4] A. McCallum, K. Nigam, L.H. Ungar, Efficient clustering of high-dimensional data sets with application to reference matching, in: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000, pp. 169–178.

[5] A. Popescul, G.W. Flake, S. Lawrence, L.H. Ungar, C.L. Giles, Clustering and Identifying Temporal Trends in Document Databases, I EEE ADL (2000) 173-182.

[6] C. Andrieu, N.D. Freitas, A. Doucet, M. Jordan, An introduction to MCMC for machine learning, Journal of Machine Learning 50 (2003) 5–43.

[7] C. Erten, P.J. Harding, S.G. Kobourov, K. Wampler, G. Yee, Exploring the Computing Literature using Temporal Graph Visualization, Technical Report, Department of Computer Science, University of Arizona, 2003.

[8] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to ad hoc information retrieval, in: Proceedings of the 24th ACM SIGIR International Conference on Research and Development in Information Retrieval, 2001, pp. 334–342.

[9] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[10] D.M. Blei, J.D. Lafferty, Dynamic topic models, in: Proceedings of the ICML, 2006, pp. 113–120.

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

[12] D. Hawking, Challenges in enterprise search, in: Proceedings of the 15th Conference on Australasian Database, 2004.

[13] D. Mimno, A. McCallum, Expertise modeling for matching papers with reviewers, in: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007, pp. 500–509.

[14] D. Petkova, W.B. Croft, Generalizing the language modeling framework for named entity retrieval, in: Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, 2007.

[15] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, ArnetMiner: extraction and mining of academic social networks, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.

[16] J. Zhang, J. Tang, L. Liu, J. Li, A mixture model for expert finding, in: T. Washio et al. (Eds.), Proceedings of the PAKDD, LNAI, vol. 5012, Springer, Heidelberg, 2008, pp. 466–478.

[17] J. Zhang, J. Tang, J. Li, Expert finding in a social network, in: R. Kotagiri, P. Radha Krishna, M. Mohania, E. Nantajeewarawat (Eds.), Proceedings of the DASFAA, LNCS, vol. 4443, Springer, Heidelberg, 2007, pp. 1066–1069.

[18] J. Zeng, S. Zhang, Variable space hidden Markov model for topic detection and analysis, Knowledge-Based Systems 21 (2008) 704–708.

[19] K. Balog, L. Azzopardi, M. de Rijke, Formal models for expert finding in enterprise corpora, in: Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, 2006, pp. 43–55.

[20] K. Balog, T. Bogers, L. Azzopardi, M. Rijke, A. Bosch, Broad expertise retrieval in sparse data environments, in: Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, 2007, pp. 551–558.

[21] L. Azzopardi, M. Girolami, K.v. Risjbergen, Investigating the relationship between language model perplexity and IR precision-recall measures, in: Proceedings of the 26th ACM SIGIR International Conference on Research and Development in Information Retrieval, Toronto, Canada, July 28–August 1, 2003.

[22] M. Li, L. Zhang, Multinomial mixture model with feature selection for text clustering, Knowledge-Based Systems 21 (2008) 704–708.

[23] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the 20th International Conference on Uncertainty in Artificial Intelligence (UAI), Canada, 2004.

[24] O. Alonso, M. Gertz, R. Baeza-Yates, On the Value of Temporal Information in Information Retrieval, in: ACM SIGIR Forum, vol. 41, December 2007.

[25] P. Mutschke, Mining networks and central entities in digital libraries: a graph theoretic approach applied to co-author networks, Intelligent Data Analysis (2003) 155–166.

[26] R. Nallapati, W. Cohen, S. Ditmore, J. Lafferty, K. Ung, Multiscale topic tomography, in: Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007.

[27] S. White, P. Smyth, Algorithms for estimating relative importance in networks, in: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp. 266–275.

[28] T. Hofmann, Probabilistic latent semantic analysis, in: Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, July 30–August 1, 1999.

[29] T.L. Griffiths, M. Steyvers, Finding scientific topics, in: Proceedings of the National Academy of Sciences (NAS), USA, 2004, pp. 5228–5235.

[30] X. Liu, J. Bollen, M.L. Nelson, H. V de Sompl, Co-authorship networks in the digital library research community, Information Processing and Management 41 (6) (2005) 681–682.

[31] X. Wang, A. McCallum, Topics over time: a non-Markov continuous-time model of topical trends, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 424–433.

[32] Y. Cao, J. Liu, S. Bao, H. Li, Research on Expert Search at Enterprise Track of TREC, 2005.

[33] Y. Li, J. Tang, Expertise search in time-varying social network, in: Proceedings of International Asia-Pacific Web Conference and Web-Age Information Management (APWeb-WAIM), 2008.

[34] Z. Nie, Y. Ma, S. Shi, J. Wen, W. Ma, Web object retrieval, in: Proceedings of the International World Wide Web (WWW), 2007, pp. 81–90.