Knowledge-Based Systems xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Using time topic modeling for semantics-based dynamic research interest finding

Ali Daud

Department of Computer Science, Sector H-10, International Islamic University, Islamabad 44000, Pakistan

ARTICLE INFO

Article history:
Received 12 December 2010
Received in revised form 26 July 2011
Accepted 26 July 2011
Available online xxxx

Keywords:
Dynamic research interests
Exchangeability of topics
Time topic modeling
Social networks
Unsupervised machine learning

ABSTRACT

Researchers interests finding has been an active area of investigation for different recommendation tasks. Previous approaches for finding researchers interests exploit writing styles and links connectivity by considering time of documents, while semantics-based intrinsic structure of words is ignored. Consequently, a topic model named Author-Topic model is proposed, which exploits semantics-based intrinsic structure of words present between the authors of research papers. It ignores simultaneous modeling of time factor which results in exchangeability of topics problem, which is important factor to deal with when finding dynamic research interests. For example, in many real world applications, like finding reviewers for papers and finding taggers in the social tagging systems one need to consider different time periods. In this paper, we present time topic modeling approach named Temporal-Author-Topic (TAT) which can simultaneously model text, researchers and time of research papers to overcome the exchangeability of topics problem. The mixture distribution over topics is influenced by both co-occurrences of words and timestamps of the research papers. Consequently, topics occurrence and their related researchers change over time, while the meaning of particular topic almost remains unchanged. Proposed approach is used to discover topically related researchers for different time periods. We also show how their interests and relationships change over a time period. Empirical results on large research papers corpus show the effectiveness of our proposed approach and dominance over Author-Topic (AT) model, by handling the exchangeability of topics problem, which enables it to obtain similar meaning of particular topic overtime.

1. Introduction

A lot of information on the Web has provided us with many challenging knowledge discovery problems, one of which is researchers' interests' discovery in academic social networks. Unfortunately, most of the existing work conducted to solve this problem ignored semantics-based structure of words, while topic modeling approaches considered semantics-based structure of words, but ignored simultaneous modeling of the time factor for different time periods. Web is highly dynamic, so the time factor cannot be ignored for most of the knowledge discovery problems these days. Most of the datasets such as research papers, tagging systems and blogs do not have static co-occurrence patterns; they are instead highly dynamic. The data are collected over different periods of time and data patterns keeps on changing, by showing rising or falling trends overtime. By finding dynamic researchers interests different recommendation tasks can be fulfilled, such as finding; reviewers for papers, project collaborators, supervisor and program committee members for conferences.

Some important scenarios about researchers' interests can be; firstly, an author A was mainly focused on biological gene networks until 2004 and published a lot of papers about this topic; afterwards he switched his concentration to image processing and not published many papers. His discovered interest in 2009 can still be biological gene networks if we ignore time factor, while this is impractical. Secondly, it is also possible that other researchers' jump into or start writing on the same topic with author A and pushes his ranking backward by publishing more papers on that specific topic. Thirdly, a researcher can be focused on more than one topic with high publishing rate at the same time. On the basis of aforementioned scenarios some intuitive questions about dynamic researchers' interests are; are the discovered researchers writing on a specific topic for each given year can be same? Are researchers' relationships for each given year can be same? Intuitively the answer to both these questions is simply "No".

In the past, efforts made to solve this problem can be divided into three major frameworks which are; (1) stylistic features (such as sentence length), author attribution and forensic linguistics to identify what the author wrote on a given piece of text [10,12] and (2) graph linkage based approaches, in which co-authorship or co-venue ship (writing for same conference or journal) based explicit connections are exploited [11,16,23] and (3) topic modeling which captures semantics-based intrinsic structure of words present between the documents [7], such as, Author-Topic (AT) model

E-mail address: ali_msdb@hotmail.com

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.07.015

[15,18,19] which considers static researchers interests or we can say that it models year by year interest individually. For example, if we have five years data, the AT model can be run for each year for finding dynamic researchers interests. Modeling researcher's interests by running model individually for each year will result in exchangeability of topics problem, which means that a topic model in different runs will not have similar topics and the order of topics will also not be the same. In general, first two types of frameworks ignored semantics-based intrinsic structure of words, while in AT individual year wise researchers interests are modeled. Later, Topics over Time (TOT) [22], a topic modeling approach was purposed to capture the evolution of topics by introducing a time node in topic model to handle the exchangeability of topics problem, but it did not consider researchers' interests. We are motivated to capture the evolution of researchers' interests with respect to topics by dependently modeling all years simultaneously.

In this paper, we combined the static researchers interests modeling idea of AT and capturing the evolution of topics idea of TOT to propose Temporal-Author-Topic (TAT) approach, which is a variation of Author-Conference-Topic model (ACT1) [20]. TAT models the dynamic research interests with respect to time without changing the meaning of topics for different years unlike AT model. Simultaneously modeling of time for all years enables TAT to handle the exchangeability of topics problems. Empirical results and discussions elaborate the importance of problem formulization and usefulness of TAT over AT.

Here it is necessary to mention that exploitation of researchers' interests (who is writing on what topic without any discrimination between renowned and not-renowned publication venues) and expert finding [6,8] (who is most skilled on what topic with the discrimination between renowned and not-renowned publication venues) are notably two different knowledge discovery problems.

The novelty of work described in this paper lies in the

(1) formalization of the dynamic researchers interests discovery problem using topic models,
(2) proposal of hybrid (TAT approach) which can handle the exchangeability of topics problem for dynamic researcher interest finding,
(3) experimental verification of the effectiveness of our approach on the real world dataset.

To the best of our knowledge, we are the first to deal with the dynamic researchers' interests' discovery problem by proposing a topic modeling approach, which can implicitly capture word–word, word–author and author–author, word–time and author–time relationships, simultaneously.

The rest of the paper is organized as follows. Section 2 provides problem formulization for dynamic researcher's interest finding. In Section 3, we introduce motivation for time author topic modeling for dynamic researchers' interests' discovery and illustrate our proposed approach with its parameters estimation details. In Section 4, corpus, parameter settings, baseline approach, with empirical studies and discussions about the results are given. Section 5 brings this paper to the conclusions and future work.

2. Finding dynamic research interests

Dynamic research interests finding focuses on discovering the right person related to a specific knowledge domain for different time periods e.g. years in this work. The question can be like "Who are the authors writing on topic Z for year Y? Instead of just what are the authors' interests on the topic Z?" In general dynamic research interests finding process, main task is to probabilistically rank discovered authors for different years. To our interest, each publication contains some title words and names which usually cover most of the highly related sub research areas of the researchers. We think that latent topics based correlations between the researchers publishing papers by simultaneously considering time effects is an appropriate way for semantics-based dynamic research interests discovery.

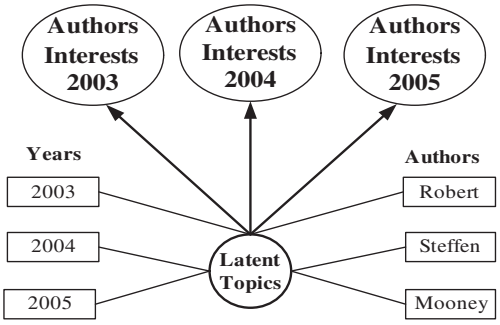

Fig. 1. Dynamic research interests finding.

We denote a document d as a vector of N_d words with a specific year y, an author a on the basis of his accepted papers, and formulize the problem as: Given a document d with N_d words having a stamp of year y, and $\mathbf{a}_c$ authors of a document d, discover related authors of a specific domain for different years. Fig. 1 provides pictorial representation of the formulization of the problem.

3. Time author topic modeling

In this section, before describing our Temporal-Author-Topic (TAT) approach, we will first briefly describe topic modeling, how documents, researchers' interests and evolution of topics are modeled.

3.1. Topic modeling

Fundamental topic modeling assumes that there is a hidden topic layer $Z = \{z_1, z_2, z_3, \ldots, z_t\}$ between the word tokens and the documents, where z_i denotes a latent topic and each document d is a vector of N_d words $\mathbf{w}_d$. A collection of D documents is defined by $D = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \ldots, \mathbf{w_d}\}$ and each word w_{id} is chosen from a vocabulary of size V. 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 word having generated probability of word given topic [4]. A hidden topic layer based approach is used for query-concept matching for digital health ecosystems service matchmaking [26], while this paper uses hidden topic layer with time to handle exchangeability of topics problem for researcher's interest finding.

3.2. Modeling documents with topics (LDA)

Latent Dirichlet Allocation (LDA) [4,13] is a 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

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

Please cite this article in press as: A. Daud, Using time topic modeling for semantics-based dynamic research interest finding, Knowl. Based Syst. (2011), doi:10.1016/j.knosys.2011.07.015

3.3. Modeling authors interests (Author model and AT model)

The Author model [14] was proposed to model documents text and its author's interests. For each document d, a set of authors' $\mathbf{a}_d$ is observed. To generate each word an author r, is uniformly sampled from the set of authors, and then a word w is generated from an author-specific multinomial distribution Φ_a over words of a document for that topic. Later, following LDA basic idea of modeling words of documents with latent topics and Author model basic idea of modeling words and authors of documents without latent topics, words and authors of documents are modeled by considering latent topics to discover the research interests of authors [19]. In AT, each author (from set of A authors) of a document d is associated with a multinomial distribution θ_a over topics is sampled from Dirichlet α and each topic is associated with a multinomial distribution Φ_z sampled from Dirichlet β over words of a document for that topic [please see Eq. (2)]. In AT time factor was not taken into account so static interests of researchers were discovered

$$P(w|a,d,\theta,\Phi)=\sum_{z=1}^{T}P(w|z,\Phi_z)P(z|a,\theta_a) \tag{2}$$

3.4. Modeling evolution of topics (DTM and TOT)

Blei and Lafferty [5] proposed dynamic topic model (DTM) which can capture the evolution of topics in a sequentially organized data. However, they ignored the natural term drift by time discretization, which can explicitly capture the rise and fall in the popularity of topics. Later, Wang and McCallum discussed time discretization limitation of DTM and proposed TOT [22]. In TOT 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 with a documents stamp y is generated from topic-specific multinomial distributions Φ_z and beta distribution Ψ_z, respectively, over words and time stamp of a document for that topic. In DTM and TOT researchers interests were not considered with the evolution of topics (see Fig. 2).

3.5. Modeling temporal authors interests with topics (Temporal-Author-Topic approach)

Firstly the basic ideas presented in AT [18,19] and TOT [22] models of modeling words and authors and words and time of documents respectively, became the intuition of modeling words, time and authors of documents together for discovering dynamic interests and relationships of researchers. Secondly, Author-Topic model can be used for finding researchers interests for each year individually, but due to exchangeability of topics problem one cannot obtain same topics for each year and the order of topics will also be different. This motivated us to introduce a time node in topic model by proposing TAT approach which can obtain same topics for each year and the order of topics is also same.

In our approach for modeling temporal interests of authors, we viewed a document as a composition of words with each word having the publishing year of document as time stamp along with its authors. Symbolically, a collection of $\mathbf{D}$ documents can be written as: $\mathbf{D}=\{(\mathbf{w}_1,\mathbf{a}_1,y_1),(\mathbf{w}_2,\mathbf{a}_2,y_2),\ldots,(\mathbf{w}_d,\mathbf{a}_d,y_d)\}$, where $\mathbf{w}_d$ is word vector chosen from a vocabulary of size V, $\mathbf{a}_d$ is author vector and y_d is the time stamp of document d.

TAT approach considers that an author is responsible for generating some latent topics of the documents on the basis of semantics-based intrinsic structures of words with time factor. In the proposed model, each author (from set of A authors) of a document d is associated with a multinomial distribution θ_a over topics and each topic is associated with a multinomial distribution Φ_z over words and multinomial distribution Ψ_z with a time stamp for each word of a document for that topic. So, θ_a, 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 of a document d is given as:

$$P(w,y|a,d,\Phi,\Psi,\theta)=\sum_{z=1}^{T}P(w|z,\Phi_z)P(y|z,\Psi_z)P(z|a,\theta_a) \tag{3}$$

The generative process of TAT is as follows:

For each author $a=1,\ldots,K$ of document d
Choose θ_a from Dirichlet (α)
For each topic $z=1,\ldots,T$
Choose Φ_z from Dirichlet (β)
Choose Ψ_z from Dirichlet (γ)
For each word $w=1,\ldots,N_d$ of document d
Choose an author a uniformly from all authors $\mathbf{a}_d$
Choose a topic z from multinomial (θ_a) conditioned on a
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 utilized [1,13] for parameter estimation in our approach, which has two latent variables z and a; the conditional posterior distribution for z and a is given by:

$$P(z_i=j,a_i=k|w_i=m,y_i=n,\mathbf{z}_{-i},\mathbf{a}_{-i},\mathbf{a}_d)\propto\frac{n_{-i,j}^{(wi)}+\beta}{n_{-i,j}^{(\cdot)}+w\beta}\frac{n_{-i,j}^{(yi)}+\gamma}{n_{-i,j}^{(\cdot)}+Y\gamma}\frac{n_{-i,j}^{(ai)}+\alpha}{n_{-i,j}^{(ai)}+A\alpha} \tag{4}$$

where $z_i=j$ and $a_i=k$ represent the assignments of the ith word in a document to a topic j and author k respectively, $w_i=m$ represents the observation that ith word is the mth word in the lexicon, $y_i=n$

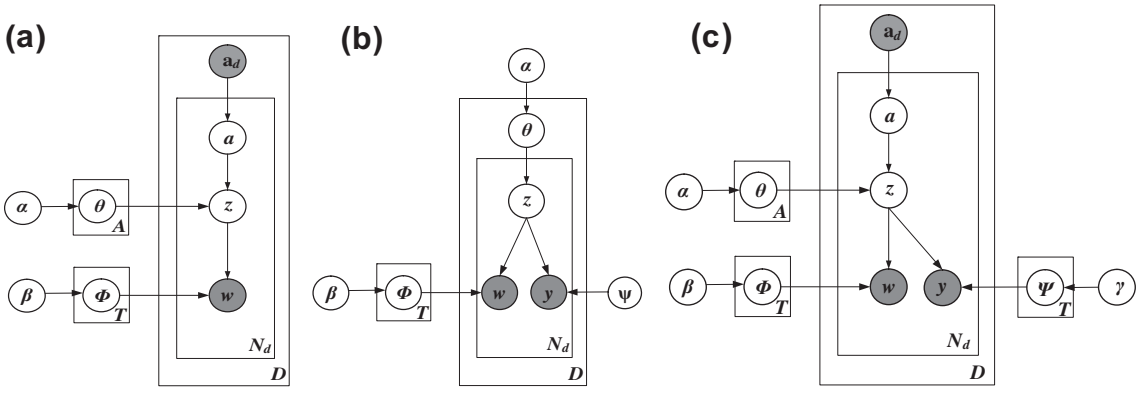

Fig. 2. TAT is shown with two inspiration models: (a) Author-Topic model (AT) [18], (b) Topics over Time (TOT) and (c) Temporal-Author-Topic Approach (TAT) [22].

Please cite this article in press as: A. Daud, Using time topic modeling for semantics-based dynamic research interest finding, Knowl. Based Syst. (2011), doi:10.1016/j.knosys.2011.07.015

Table 1
Generative summary of TAT and related models.

Model	Summarized generative process and problem solved
AT	An author of a document is responsible for generating words for documents on the basis of latent topics. *Static Authors Interests Discovery*
TOT	A document is responsible for generating words with time stamps for documents on the basis of latent topics. *Evolution (Dynamism) of Topics*
ACT1	An author of a document is responsible for generating words with conference stamp for documents on the basis of latent topics. *Static Conferences and Authors Discovery*
TAT	An author of a document is responsible for generating words with time stamps for documents on the basis of latent topics. *Temporal (Dynamic) Authors Interests Discovery*

represents ith year of paper publishing, attached with the nth word in the lexicon and z_{-i} and a_{-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 $n_{-i,j}^{(ai)}$ and 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 A is the number of authors. "." 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 $W \times T$ (word by topic), $Y \times T$ (year by topic) and $T \times A$ (topic by author) count matrices. From these count matrices, topic-word distribution Φ, topic-year distribution Ψ and Author-Topic distribution θ can be calculated as:

$$P(w|z) = \Phi_{zw} = \frac{n_{-i,j}^{(wi)} + \beta}{n_{-i,j}^{(\cdot)} + w\beta} \tag{5}$$

$$P(y|z) = \Psi_{zy} = \frac{n_{-i,j}^{(yi)} + \gamma}{n_{-i,j}^{(\cdot)} + Y\gamma} \tag{6}$$

$$P(z|a) = \theta_{az} = \frac{n_{-i,j}^{(ai)} + \alpha}{n_{-i,\cdot}^{(ai)} + A\alpha} \tag{7}$$

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

Now finally by using joint conditional probability, we can obtain the probability of an author a given topic z and year y as:

$$P(a|z,y) = \frac{P(z,y|a) \cdot P(a)}{P(z,y)}, \quad \text{where}$$
$$P(z,y|a) = P(z|a) \cdot P(y|a) \quad \text{and} \quad P(y|a) = \sum_z P(y|z) \cdot P(z|a) \tag{8}$$

Here, for calculating $P(a)$ we simply used the number of publications of one author in a year. For more simplicity some works assume it uniform [3] and Propagation approach can also be used to calculate it in a more complex way [25]. For better understanding of difference between proposed approach and related models, Table 1 provides the general description of models and problems handled by using these models.

4. Experiments

4.1. Corpus

We downloaded five years (2003–2007) research publications corpus of conferences from DBLP [9]. In total, we extracted 112,317 authors, 90,124 publications for 261 conferences. We then processed corpus by (a) removing stop-words, punctuations and numbers (b) down-casing the obtained words of publications, and (c) removing words and authors that appear less than three times in the corpus. This led to a vocabulary size of $V = 10{,}872$, a total of 572,592 words and 26,078 authors in the corpus. Fig. 3 shows fairly smooth yearly data distribution for number of publications (D) and authors (A) in conferences.

There is certainly some noise in data of this form especially author names which were extracted automatically by DBLP from PDF, postscript or other document formats. For example, for some very common names there can be multiple authors (e.g. L Ding or J Smith). This is a known limitation of working with this type of data (please see [17] for details). There are algorithmic techniques for name disambiguation that could be used to automatically solve these kinds of problems; however, in this work we do not focus on name disambiguation problems.

4.2. Parameter settings

In our experiments, for 150 topics Z the hyper-parameters α, β and γ were set at $50/Z$, 0.01, and 0.1. Topics are set at 150 on the basis of human judgment of meaningful topic plus measured perplexity [4], a standard measure for estimating the performance of probabilistic models with the lower the best, for the estimated topic models. Teh et al. [21] proposed a solution for automatic selection of number of topics, which can also be used for topic optimization, but we are not focused on that in this work. All experiments were carried out on a machine running Windows XP 2006 with Intel (R) Core (TM) 2 Duo CPU T5670 (1.80 GHz) and 2 GB memory.

4.3. Baseline approach

We attempted to qualitatively compare TAT with AT and used same number of topics for evaluation. Dataset was portioned by year and for each year all the words and authors were assigned to their most likely topics using AT model. The number of Gibbs sampler iterations used for AT is 1000 and parameter values same as the values used in [18].

4.4. Results and discussion

4.4.1. Topically related authors for different years

We discovered and probabilistically ranked researchers related to a specific area of research on the basis of latent topics for different years. Table 2 illustrates 4 different topics out of 150, discovered from the 1000th iteration of a particular Gibbs sampler run. The words associated with each topic are quite intuitive and precise in the sense of conveying a semantic summary of a specific

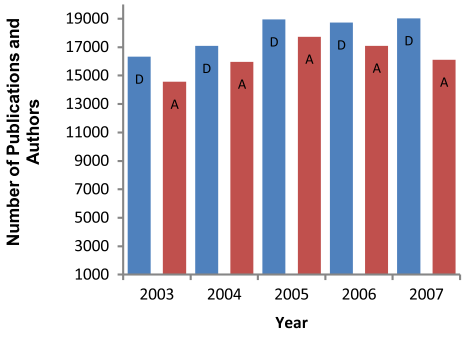

Fig. 3. Histogram illustrating data distribution.

area of research. The authors associated with each topic for different years are quite representative. Here it is necessary to mention that top 10 authors associated with the topics for different years are not the experts of their fields, instead are the authors who produced most words for that topic in a specific year. For example, Baowen Xu is known for software engineering, by analyzing DBLP data we have found that he has published papers having high probability words related to Data Mining (DM) topic during the years he is related to that topic, while he is not related to DM topic later because of not producing high probability words for that topic.

For Data Mining topic Jiawei Han, XML Databases (XMLDB) topic Surajit Chaudhuri, and for Bayesian Learning (BL) topic Andrew Y. Ng has been leading authors for different years mostly. While, other authors related to different topics for different years kept on changing their ranking due to writing less papers on the topics or other authors writing more on the same topic.

In addition, by doing analysis of researchers' home pages and DBLP data, we found that all highly ranked authors for different years have published papers on their assigned topics for specific topics; "*no matter where they are publishing and they are old or new researchers*". For example, Jianhua Feng (new researcher) for XMLDB topic started writing on this topic after 2004 and then has published many papers in the following years especially in 2006 (ranked first) and 2007 (ranked second). He published most papers in WAIM (country level conference), DASFAA (continent level conference) and some in WWW (world level conference) and here we considered world level conference among the best (world class conference) in that research area. While, other top ranked authors for this topic Surajit Chaudhuri (old researcher) in 2006 (ranked second) and 2007 (ranked first) published many papers in SIGMOD (world level conference), ICDE (world level conferences) and VLDB (world level conference) and produced many words for XMLDB topic. He is continuously publishing over the years for this topic. Here, Jianhua Feng and Surajit Chaudhuri produced most words for this topic and ranked higher without the discrimination of where they published and from when they are publishing. This matches well with the statement stated above and provides qualitative supporting evidence for the effectiveness of the proposed approach.

A direct comparison with the previous approaches is not fair in terms of perplexity [4], as previous topic modeling approaches were unable to discover dynamic researchers' interests with considering time factor. To measure the performance in terms of precision and recall [2] is also out of question due to unavailability of standard dataset and use of human judgments cannot provide appropriate (unbiased) answers for evaluating dynamic researchers' interests finding methods. So, we compared TAT approach with AT [18], TAT approach can have same meaning for particular topic overtime, but by ignoring time factor AT model changed the meaning of particular topic overtime (inability to discover similar topics for different years or exchangeability of topics problem). It concludes that approaches which did not consider time factor are unable to discover approximately similar topics for different years. We can say that the time-based solution provided by us is well justified and produced quite promising and functional results.

4.4.2. Exchangeability of topics effect on different years authors

AT model does not consider time information simultaneously 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 is not theoretically considered to be similar to topic z_i in the other runs of the algorithm [7].

Consequently, when we ran AT for finding research interests for five years individually that resulted into three main problems. Firstly, the topics numbers were not similar for different years. Secondly, the probabilistically related words were also not exactly leading to same area of interest. Thirdly, the authors related to a topic for different years are very diverse and not accurate. The problems result in having topically related biased researchers for different topics.

We show "Data Mining" and "Support Vector Machines" topics and their related authors for AT model. Here, data mining is a general topic while support vector machine is a bit specific topic. We see in Table 3 that the words for data mining topic and the authors found for different years are not very much different. While support vector machine topic words for different years obtained by AT explains problem of not having similar topic number and probabilistic words for each because of modeling for each year's independently. For example, topic for year 2003 has words programming, abstract, gap, demonstration and other words which are not in other year topics, consequently the authors producing these words are not similar as words in other topics, which will result in incorrect authors for support vector machines topic, same will happen for other specific topics also. We see that Thorsten Joachims who proposed support vector machine is just found for the year 2003, while in Table 2 our proposed method found his interests in this topic every year, which matches with real world data as by analyzing his home page we have found that he is continuously publishing related to this topic. Table 2 also shows that our proposed method found that Bing Li, Ravishankar K. Iyer, Bin Wang and Mahmood S. Karnal have continues interest in support vector machine topic which is supported by papers published by them on this specific topic for different years, while AT model was unable to find these authors related to this topic.

We carefully analyzed the results for support vector machines topic words and found that the words except top three or four are different for each year for AT, especially the words for year 2007 are not even having top three or four words similar to previous year's topic words. Infect the topic for year 2007 even do not have "machines" word and does not even look like support vector machines topic. Same as different words the authors found by AT for different years are also different as there is no author who has the same interest for these years for support vector machines topic, which does not match with the real world data. It is simply not possible that for one topic each year authors writing on that topic with high frequency are different. Conclusively, one can say that when some topic is general like "Data Mining" the approach like AT which model each year independently can work fine, but it is unable to perform well for specific topics like support vector machines, consequently we need a method which can model years at once to find more precise topics and interests of authors over the years.

4.4.3. Temporal social network of researcher

TAT approach can also be used for dynamic correlation discovery between authors for different years, as compared to only discovering static authors' correlations [18]. To illustrate how it can be used in this respect, distance between authors i and j is calculated by using Eq. (9) for Author-Topic distribution for different years

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

We calculated the dissimilarity between authors; smaller dissimilarity value means higher correlation between the authors. Table 4 shows topically related authors with Jiawei Han for different years. Here, it is obligatory to mention that top 10 authors related to Jiawei Han are not the authors who have co-authored with him mostly, but rather are the authors that tend to produce most words

Please cite this article in press as: A. Daud, Using time topic modeling for semantics-based dynamic research interest finding, Knowl. Based Syst. (2011), doi:10.1016/j.knosys.2011.07.015

Table 2
An illustration of 4 discovered topics from 150 topics. Each topic is shown with the top 10 words (first column) and authors that have highest probability conditioned on that topic for each year (second to sixth column). Titles are our interpretation of the topics.

Words	Prob.	Authors	Prob.	Authors	Prob.	Authors	Prob.	Authors	Prob.	Authors	Prob.
Data Mining (DM)											
Topic 142		Year 2003		Year 2004		Year 2005		Year 2006		Year 2007	
Mining	0.20871	Jiawei Han	0.2	Jiawei Han	0.2	Jiawei Han	0.2	Jiawei Han	0.2	Jiawei Han	0.2
Patterns	0.07798	Francesco Bonchi	0.0845	Hui Xiong	0.1023	Francesco Bonchi	0.0646	Hui Xiong	0.0561	Hui Xiong	0.0866
Rules	0.05193	Baowen Xu	0.0732	George Karypis	0.0683	Srinivasan Parth.	0.0634	Francesco Bonchi	0.0439	*Jian Pei*	0.0311
Frequent	0.04291	Hui Xiong	0.0539	Reda Alhajj	0.0573	Baowen Xu	0.0509	Srinivasan Parth.	0.0431	Christopher Kruegel	0.0264
Pattern	0.04155	George Karypis	0.0525	Francesco Bonchi	0.0515	*Jian Pei*	0.0458	Reda Alhajj	0.0315	Reda Alhajj	0.0243
Association	0.04121	*Jian Pei*	0.0507	Srinivasan Parth.	0.0385	Reda Alhajj	0.0424	Olfa Nasraoui	0.0266	Martin Ester	0.0237
Discovery	0.023	Srinivasan Parth.	0.0474	Shiwei Tang	0.0385	Shiwei Tang	0.0414	*Jian Pei*	0.0264	Francesco Bonchi	0.0203
Databases	0.02283	Takeaki Uno	0.0375	Baowen Xu	0.0324	George Karypis	0.0321	Martin Ester	0.0256	Olfa Nasraoui	0.0197
rule	0.01908	Jeffrey Xu Yu	0.0341	Jianyong Wang	0.0297	Hui Xiong	0.0275	Shiwei Tang	0.0246	Won Suk Lee	0.0197
Discovering	0.01619	Won Suk Lee	0.0327	*Jian Pei*	0.0262	Ke Wang	0.0266	Jianyong Wang	0.019	Takeaki Uno	0.0181
Support Vector Machines (SVM)											
Topic 18		Year 2003		Year 2004		Year 2005		Year 2006		Year 2007	
Support	0.21272	Ravishankar K. Iyer	0.2	Ravishankar K. Iyer	0.2	Thorsten Joachims	0.2	Bing Li	0.2	Thorsten Joachims	0.2
Vector	0.08597	Thorsten Joachims	0.1341	George Karypis	0.1861	Ravishankar K. Iyer	0.0829	Thorsten Joachims	0.1239	Bin Wang	0.1383
Machines	0.06282	Bing Li	0.1269	Laurie A. Williams	0.1562	Bing Li	0.0746	P. Madhusudan	0.0804	Mohamed S. Kamel	0.1276
Machine	0.03852	Bin Wang	0.1236	Thorsten Joachims	0.1341	Onur Mutlu	0.074	George Karypis	0.0586	Thomas Baumlck	0.1212
Regression	0.01074	George Karypis	0.1163	Bing Li	0.1038	Bin Wang	0.0615	Mohamed S. Kamel	0.0529	Massimo Melucci	0.1019
Kernel	0.00755	P. Madhusudan	0.1033	P. Madhusudan	0.0826	P. Madhusudan	0.0514	Manuel V. Hermenegildo	0.0518	Ravishankar K. Iyer	0.0995
Complex	0.00379	Ioannis Tsamardinos	0.0967	Manuel V. Hermenegildo	0.0627	George Karypis	0.0463	Bin Wang	0.0487	Kamesh Madduri	0.083
Svms	0.00321	Matthew B. Dwyer	0.0736	Lex Wolters	0.0607	Jun Yan	0.0349	Alexander Romanovsky	0.046	Bing Li	0.0688
Multiresolution	0.00321	Lawrence O. Hall	0.0693	Matthew B. Dwyer	0.0589	Jun Li	0.0336	Matthew B. Dwyer	0.0441	Jie Zhang	0.0686
High	0.00292	B. Earl Wells	0.0639	Palash Sarkar	0.0527	Sudarshan K. Srinivasan	0.0321	Jun Yan	0.0436	Byron Cook	0.0638
Bayesian Learning (BL)											
Topic 111		Year 2003		Year 2004		Year 2005		Year 2006		Year 2007	
Learning	0.21209	Andrew Y. Ng	0.2	Andrew Y. Ng	0.2	Andrew Y. Ng	0.2	Andrew Y. Ng	0.2	Andrew Y. Ng	0.2
Bayesian	0.04646	Michael I. Jordan	0.1635	Arindam Banerjee	0.1753	Ling Li	0.0606	Tao Li	0.0763	Arindam Banerjee	0.1963
Inference	0.02233	Tao Li	0.1574	Alexey Tsymbal	0.0753	Tao Li	0.0573	Alexey Tsymbal	0.0732	S. V. N. Vishwanathan	0.1599
Classifiers	0.02167	Ling Li	0.1514	Michael I. Jordan	0.0701	Zoubin Ghahramani	0.0495	Harry Zhang	0.0521	Tao Li	0.1546
Semi	0.02068	Dale Schuurmans	0.1372	Tao Li	0.0675	Harry Zhang	0.0447	Ling Li	0.0471	Ling Li	0.1393
Classification	0.02068	Alexey Tsymbal	0.1318	Bernhard Scholkopf	0.0529	S.V.N. Vishwanathan	0.0444	S.V.N. Vishwanathan	0.0444	Xiaofei He	0.084
Supervised	0.02051	Naftali Tishby	0.1269	Ling Li	0.0519	Volker Tresp	0.0414	Dale Schuurmans	0.04	Jie Hu	0.0808
Reinforcement	0.02002	Daphne Koller	0.119	Xiaofei He	0.0514	Xiaofei He	0.04	Rohit Singh	0.0379	Bernhard Scholkopf	0.0802
probabilistic	0.01671	Rocco A. Servedio	0.1173	Zoubin Ghahramani	0.0495	Ira Cohen	0.0349	Zoubin Ghahramani	0.0366	Terran Lane	0.0794
Models	0.01489	Volker Tresp	0.1035	Csaba Szepesuari	0.0484	Rocco A. Servedio	0.0342	Volker Tresp	0.0345	Avi Pfeffer	0.0704
XML Databases (XMLDB)											
Topic 18		Year 2003		Year 2004		Year 2005		Year 2006		Year 2007	
XML	0.16606	Surajit Chaudhuri	0.2	Surajit Chaudhuri	0.2	Surajit Chaudhuri	0.2	Jianhua Feng	0.2	Surajit Chaudhuri	0.2
Query	0.08406	Divesh Srivastava	0.1038	Jayant R. Haritsa	0.1161	Philip A. Bernstein	0.106	Surajit Chaudhuri	0.1239	Jianhua Feng	0.157
Database	0.05458	Raymond K. Wong	0.0825	Carlos A. Heuser	0.0727	Kevin Chen-Chuan	0.0961	Raymond K. Wong	0.0804	Kevin Chen-Chuan	0.0961
Databases	0.05113	Jayant R. Haritsa	0.0774	Tok Wang Ling	0.0719	Dan Suciu	0.0912	Dimitri Theodoratos	0.0586	Christoph Koch	0.0959
Relational	0.04016	Dan Suciu	0.0676	Raymond K. Wong	0.0675	Tok Wang Ling	0.0847	Raghu Ramakrishnan	0.0529	Dan Suciu	0.0946
Queries	0.03467	Christoph Koch	0.06	Kevin Chen-Chuan	0.0437	Donald Kossmann	0.0811	Sourav S. Bhowmick	0.0518	Donald Kossmann	0.0721
Schema	0.02981	Carlos A. Heuser	0.0533	Raghu Ramakrishnan	0.0427	Dimitri Theodoratos	0.0788	Divesh Srivastava	0.0487	Carlos A. Heuser	0.0711
Querying	0.02636	Hongjun Lu	0.0486	Sourav S. Bhowmick	0.0418	Jianhua Feng	0.0757	Christian S. Jensen	0.046	Divesh Srivastava	0.0601
documents	0.02401	Elke A. Rund.	0.0414	Christoph Koch	0.0392	Jayant R. Haritsa	0.0726	Erik Wilde	0.0441	Dimitri Theodoratos	0.0526
Views	0.02338	Jixue Liu	0.0402	Divesh Srivastava	0.0358	Divesh Srivastava	0.0615	Katsumi Tanaka	0.0436	Erik Wilde	0.0494

Please cite this article in press as: A. Daud, Using time topic modeling for semantics-based dynamic research interest finding, Knowl. Based Syst. (2011), doi:10.1016/j.knosys.2011.07.015

Please cite this article in press as: A. Daud, Using time topic modeling for semantics-based dynamic research interest finding, Knowl. Based Syst. (2011), doi:10.1016/j.knosys.2011.07.015

Table 3
Exchangeability of topics problem.

"Topic 33" 2003		"Topic 75" 2004		"Topic 124" 2005		"Topic 7" 2006		"Topic 18" 2007	
Words	Prob.	Words	Prob.	Words	Prob.	Words	Prob.	Words	Prob.
mining	0.121387	data	0.169117	data	0.167838	data	0.166476	mining	0.121004
data	0.118838	mining	0.099283	mining	0.119303	mining	0.122461	data	0.050026
patterns	0.039307	clustering	0.046908	patterns	0.030092	clustering	0.051855	patterns	0.036398
rules	0.033699	patterns	0.031287	streams	0.029629	streams	0.029847	privacy	0.035262
clustering	0.030130	databases	0.028990	frequent	0.021771	patterns	0.028472	pattern	0.029016
association	0.030130	frequent	0.025315	pattern	0.020385	frequent	0.016093	preserving	0.024473
frequent	0.024522	rules	0.023936	association	0.018536	databases	0.014717	frequent	0.013117
streams	0.017894	streams	0.023018	rules	0.017149	gene	0.012883	streams	0.009142
pattern	0.011777	association	0.020720	stream	0.014376	stream	0.012883	association	0.006871
sequential	0.010247	discovery	0.011991	approach	0.008366	series	0.007840	sequential	0.006303
Authors	Prob.	Authors	Prob.	Authors	Prob.	Authors	Prob.	Authors	Prob.
Jiawei Han	0.004834	Reda Alhajj	0.003317	Philip S. Yu	0.003075	Jiawei Han	0.005417	Jiawei Han	0.005593
Srinivasan Parth.	0.003021	Jiawei Han	0.002871	Haixun Wang	0.002249	Philip S. Yu	0.003580	Wei Wang	0.002586
Ming-Syan Chen	0.002527	Philip S. Yu	0.002723	Jiawei Han	0.002111	Eamonn J. Keogh	0.002308	Ling Liu	0.001794
Osmar R. Zaine	0.002527	Wei Wang	0.002277	Christos Faloutsos	0.001698	Ming-Syan Chen	0.001884	Justin Z. Zhan	0.001636
Jian Pei	0.002197	Xindong Wu	0.002129	Reda Alhajj	0.001560	Anthony K.H. Tung	0.001743	Wenliang Du	0.001477
Kotagiri Rama	0.002197	Jesacute S. A. Ruiz	0.002129	Geoffrey I. Webb	0.001423	Lizhu Zhou	0.001743	Philip S. Yu	0.001161
Sharma Chak	0.002033	Taneli Miel.	0.001980	Srinivasan Parth	0.001423	Wei Wang	0.001743	Christos Faloutsos	0.001161
Wei Wang	0.002033	Jian Pei	0.001980	Jian Pei	0.001423	Charu C. Aggarwal	0.001743	Sandra de Amo	0.001003
Raj P. Gopalan	0.002033	David Taniar	0.001832	Graham Cormode	0.001285	Srinivasan Parth.	0.001743	Jae Soo Yoo	0.001003
Jean-Franois Boul	0.002033	Michael K. Ng	0.001832	Bing Liu	0.001285	James Bailey	0.001601	Sang-Wook Kim	0.000844
"Topic 119" 2003		"Topic 27" 2004		"Topic 41" 2005		"Topic 148" 2006		"Topic 11" 2007	
Words	Prob.	Words	Prob.	Words	Prob.	Words	Prob.	Words	Prob.
support	0.099503	support	0.074217	support	0.105889	support	0.059635	analysis	0.042635
vector	0.038317	vector	0.038542	vector	0.038846	vector	0.030855	static	0.021003
machines	0.024551	machines	0.028749	machines	0.034245	machine	0.030185	automated	0.014723
programming	0.009254	machine	0.023853	order	0.006639	classification	0.024831	support	0.011932
abstract	0.004665	hierarchical	0.023153	pattern	0.005981	machines	0.010776	dynamic	0.008443
gap	0.003901	systems	0.006365	multiple	0.005324	applications	0.010776	online	0.008443
requirements	0.003901	free	0.006365	game	0.005324	fault	0.009437	algorithms	0.007745
demonstration	0.003901	study	0.005666	software	0.005324	multiple	0.006091	integrating	0.007745
visualization	0.003901	efficient	0.004966	error	0.004667	statistical	0.006091	vector	0.007048
integrated	0.003136	programming	0.004966	time	0.004667	neural	0.006091	specification	0.007048
Authors	Prob.	Authors	Prob.	Authors	Prob.	Authors	Prob.	Authors	Prob.
Mitja Lenic	0.000985	Thomas Hofmann	0.001559	Andre Carlos	0.001107	Hong Peng	0.000991	Sarfraz Khurshid	0.001224
Jun Li	0.000800	Jennifer G. Dy	0.001058	Binh Pham	0.000956	Jianna Zhang	0.000835	Peng Li	0.000724
Amund Tveit	0.000800	Leacuteon Bottou	0.000891	Chih-Jen Lin	0.000956	Keivan Kianmehr	0.000835	Anupam Basu	0.000724
Hyunjung Shin	0.000800	Muhammad Shaaban	0.000891	Antocircio de P. Braga	0.000956	Angel Fernando	0.000678	Gerth Stlash	0.000557
Hirotaka Nakayama	0.000800	Leslie Carr	0.000724	Thorsten Joachims	0.000805	Yukun Bao	0.000678	Colm O'Riordan	0.000557
Klaus Obermayer	0.000800	Toshihisa Takagi	0.000724	Atsuo Hazeyama	0.000805	Dale Miller	0.000522	Paul G. Spirakis	0.000557
Daniel P. Miranker	0.000616	Alan F. Smeaton	0.000724	Kousha Etessami	0.000654	Mao Ye	0.000522	Rasit Onur Topaloglu	0.000557
Bernard Manderick	0.000616	Alvin T. S. Chan	0.000724	Hong Hu	0.000654	Gang Chen	0.000522	Tao Xie	0.000557
Paul McNamee	0.000616	Ravishankar K. Iyer	0.000724	Hong Peng	0.000654	Max Chacoacun	0.000522	Giuliano Antoniol	0.000390
Hwanjo Yu	0.000616	Chenn-Jung Huang	0.000724	Tu Bao Ho	0.000654	Tao Li	0.000522	Wenping Wang	0.000390

Table 4
Top 10 associated authors with Jiawei Han for different years.

Jiawei Han									
2003		2004		2005		2006		2007	
Jian Pei	0.3180	Jian Pei	0.3279	Jian Pei	0.3575	Jian Pei	0.3435	Jian Pei	0.3183
Jeffrey Xu Yu	0.3271	Srinivasan Parthasarathy	0.3916	Francesco Bonchi	0.4648	Srinivasan Parthasarathy	0.4015	Won Suk Lee	0.3598
Francesco Bonchi	0.3781	Francesco Bonchi	0.4057	Srinivasan Parthasarathy	0.4725	Olfa Nasraoui	0.4305	Olfa Nasraoui	0.3824
Srinivasan Parthasarathy	0.4168	Olfa Nasraoui	0.4273	Olfa Nasraoui	0.5465	Hui Xiong	0.4329	Hui Xiong	0.4561
Hui Xiong	0.4567	Anthony K.H. Tung	0.4422	Haixun Wang	0.5632	Francesco Bonchi	0.4549	Martin Ester	0.4654
Haixun Wang	0.4786	Gao Cong	0.4608	Gang Chen	0.5753	Martin Ester	0.4661	Gang Chen	0.5005
Won Suk Lee	0.4807	Joshua Zhexue Huang	0.4616	Philip S. Yu	0.5866	Gang Chen	0.4993	Agma J. M	0.5137
Kuniaki Uehara	0.4844	Martin Ester	0.4666	Hui Xiong	0.6060	Agma J.M. Traina	0.5119	Zhoujun Li	0.5354
Olfa Nasraoui	0.4956	Show-Jane Yen	0.4697	S. Muthukrishnan	0.6148	Mete Celik	0.5184	Haixun Wang	0.5480
Takeaki Uno	0.5170	Wynne Hsu	0.4730	Jeffrey Xu Yu	0.6215	Efim B. Kinber	0.5281	Takeaki Uno	0.5531

for the same topics with him. Again the results are quite promising and realistic as most of the authors related to Jiawei Han for different years are also related to DM topic.

In addition to show the effectiveness of proposed approach for temporal relationship discovery, we calculated Symmetric KL divergence between pairs Jiawei Han and Jian Pei (Same topic and also co-authors), Jiawei Han and Francesco Bonchi (same topic, not co-authors and Bonchi ranked usually higher than Alhajj for DM topic), Jiawei Han and Reda Alhajj (same topic, not co-author and Alhajj ranked usually lower than Bonchi for DM topic), Jiawei Han and Andrew Y. Ng (different topic and not co-authors as Ng belongs to BL topic). From Table 5 we can see that Jiawei Han and Jian Pei have smallest distance for each year because they have written and co-authored on the same topic continuously. Jiawei Han and Francesco Bonchi have more distance with Jiawei Han than Jian Pei as he has written on the same topic but not co-authored with Jiawei Han. This shows that proposed approach can successfully use co-authorship information and matches well with the results presented in Table 4 for discussed authors.

Reda Alajj has more distance with Jiawei Han than Jian Pei and Francesco Bonchi, which matches well with his lower ranking in DM topic. Andrew Y. Ng has more distance than all authors shown in Table 2 related to DM because his main interest area is BL. This matches well with the results presented in Table 2.

Comparatively, AT [18] and TOT [22] are unable to discover temporal social network of researchers, as AT has not considered time information for all years simultaneously making it face exchangeability of topics problem, while TOT did not used researchers information.

4.4.4. Dynamic research interests

Now by using TAT we will show topic-wise and author-wise dynamic research interests. In Fig. 4, for DM topic Jian Pei has a stable publishing interest shows his consistency to retain his position, while Franscesco Bonchi and George Karypis either started writing less related to this topic or some other authors have influenced their interests by writing more on the same topic.

In BL topic the interest of Michael I. Jordan is temporally decreased a bit, by analyzing DBLP data we found that his number of publications decreased from 27, 21, 10, 11, 4 in order of years 2003–2007. It became the reason of producing fewer words for BL topic (his major interest topic). Ling Li almost have a stable interest for this topic, while Tao Li has a parabolic interest for the topic as by analyzing DBLP data we found he is focused on many research areas at the same time, so unable to retain his position over the years for this topic.

In Fig. 5, Reda Alhajj is a good representative of the scenario, that one researcher's interests can be focused on more than one topic with high publishing rate. He has published on DM, XMLDB and BL topics simultaneously with a little bit more focus on DM topic.

While on the other hand, Thorsten Joachim's has totally different kind of interest patterns as he is pioneer of support vector machines (SVM) and still strongly publishing related to that topic (shows clearly the importance of temporal authors interests discovery problem and effectiveness of proposed approach by matching well with the real world situation). For second and third related interests' topics clustering and semantic information retrieval he has published very little, by analyzing his publications in DBLP data we found that he used SVM as a tool to perform clustering and information retrieval tasks. Comparatively, AT [18] and TOT [22] are unable to discover topic wise dynamic researchers' interests, as AT has not considered time information for all years together making it face exchangeability of topics problem, while TOT did not used authors information.

Please cite this article in press as: A. Daud, Using time topic modeling for semantics-based dynamic research interest finding, Knowl. Based Syst. (2011), doi:10.1016/j.knosys.2011.07.015

Table 5
Symmetric KL divergence for pairs of authors.

Jiawei Han	2003	Co-auth.	2004	Co-auth.	2005	Co-auth.	2006	Co-auth.	2007	Co-auth.
Jian Pei	0.1125	1	0.0914	6	0.1203	2	0.1015	3	0.1205	1
Francesco Bonchi	0.1664	0	0.1614	0	0.1930	0	0.1583	0	0.2573	0
Reda Alhajj	0.2292	0	0.2074	0	0.2533	0	0.2954	0	0.3025	0
Andrew Y. Ng	0.2543	0	0.2493	0	0.3352	0	0.3745	0	0.3591	0

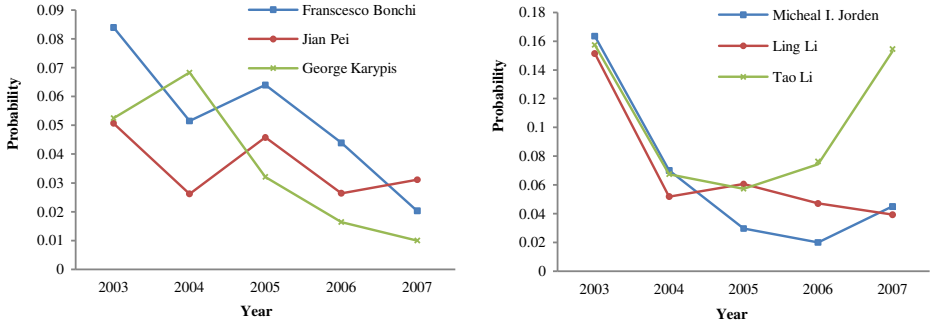

Fig. 4. Topic-wise research interests for data mining (left) and Bayesian learning (right).

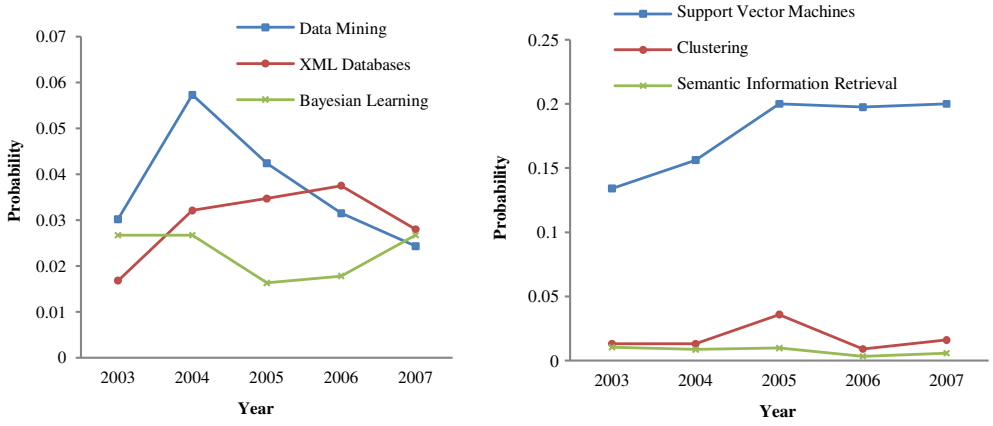

Fig. 5. Author-wise interests of Reda Alhajj (left) and Thorsten Joachims (right).

5. Conclusions and future work

This work is conducted to deal with the problem of discovering dynamics researchers' interests through modeling documents, authors and time simultaneously to handle exchangeability of topics problem. Initially discussed motivation for dynamic researchers interests modeling is well justified, as it is significant to use text, authors and time information of documents, simultaneously. Introduced TAT approach can discover and probabilistically rank researchers related to specific knowledge domains for different time periods. Dynamic semantics-based social network shown for researchers' on the basis of latent semantics is quite realistic. Dynamic researchers' interests shown matches well with the real data. TAT can handle the problem of AT model of change in the meaning of topic overtime successfully. Empirical results and discussions prove the effectiveness of proposed approach. From generic point of view, our approach can also be applied to blogs dataset for bloggers interests' discovery, news dataset for discovering news reporters' interests and active news issues and decisively any dataset which has time series text information with the authors. In future, time discretization by year problem of TAT because of using discrete probability distribution will be tried to handle by using continuous probability distribution; such as continuous time dynamic topic models used continuous stochastic processes for capturing topics dynamics [24].

Acknowledgements

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

References

[1] C. Andrieu, N.D. Freitas, A. Doucet, M.I. Jordan, An introduction to MCMC for machine learning, Journal of Machine Learning 50 (2003) 5–43.
[2] 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 Conference on Research and Development in Information Retrieval, Toronto, Canada, July 28–August 1, 2003.
[3] K. Balog, T. Bogers, L. Azzopardi, M.D. Rijke, A.V.D. Bosch, Broad expertise retrieval in sparse data environments, in: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 2007, pp. 551–558.
[4] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.
[5] D.M. Blei, and J.D. Lafferty, Dynamic topic models, in: Proceedings of the International Conference on Machine Learning (ICML), 2006, pp. 113–120.
[6] A. Daud, J. Li, L. Zhou, F. Muhammad, A generalized topic modeling approach for maven search, in: Proceedings of the International Asia–Pacific Web Conference and Web-Age Information Management (APWEB-WAIM), Suzhou, China, 2009.

[7] 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) 4 (2) (2010) 280–301.
[8] A. Daud, J. Li, L. Zhu, F. Muhammad, Temporal expert finding through generalized time topic modeling, Knowledge Based Systems (KBS) 23 (6) (2010) 615–625.
[9] DBLP Bibliography Database. <http://www.informatik.uni-trier.de/~ley/db/>.
[10] J. Diederich, J. Kindermann, E. Leopold, G. Paass, Authorship attribution with support vector machines, Applied Intelligence 19 (1) (2003).
[11] 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.
[12] A. Gray, P. Sallis, S. MacDonell, Softwareforensics: extending authorship analysis techniques to computer programs, in: Proceedings of the 3rd IAFL, Durham, NC, 1997.
[13] T.L. Griffiths, M. Steyvers, Finding scientific topics, in: Proceedings of the National Academy of Sciences (NAS), USA, 2004, pp. 5228–5235.
[14] A. McCallum, Multi-label text classification with a mixture model trained by EM, in: Proceedings of the AAAI'99 Workshop on Text Learning, 1999.
[15] 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.
[16] 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.
[17] M.E.J. Newman, Scientific collaboration networks: I. Network construction and fundamental results, Physical Review E 64 (2001) 016131.
[18] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, M. Steyvers, Learning author-topic models from text corpora, ACM Transactions on Information Systems (2009) 1–38.
[19] M. Steyvers, P. Smyth, T. Griffiths, Probabilistic author-topic models for information discovery, in: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, 2004.
[20] 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.
[21] Y.W. Teh, M.I. Jordan, M.J. Beal, D.M. Blei, Hierarhical Dirichlet Processes. Technical Report 653, Department of Statistics, UC Berkeley, 2004.
[22] 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.
[23] 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.
[24] C. Wang, M.D. Blei, D. Heckerman, Continuous time dynamic topic models, in: Proceedings of the Uncertainty in Artificial Intelligence (UAI), Helsinki, Finland, July 9–12, 2008.
[25] J. Zhang, J. Tang, J. Li, Expert finding in a social network, in: Proceedings of the International Conference on Database Systems for Advanced Applications (DASFAA), 2007.
[26] H. Dong, F.K. Hussain, Semantic service matchmaking for digital health ecosystems, Knowledge Based Systems (KBS) 24 (2011) 761–774.

Please cite this article in press as: A. Daud, Using time topic modeling for semantics-based dynamic research interest finding, Knowl. Based Syst. (2011), doi:10.1016/j.knosys.2011.07.015