A Generalized Topic Modeling Approach for Maven Search

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

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

Abstract. This paper addresses the problem of semantics-based maven search in research community, which means identifying a person with some given expertise. Traditional approaches either ignored semantic knowledge or temporal information, resulting in some right mavens that cannot be effectively identified because of non-occurrence of keywords and un-exploitation of time effects. In this paper, we propose a novel semantics and temporal information based maven search (STMS) approach to discover latent topics (semantically related soft clusters of words) between the authors, venues (conferences or journals) and time simultaneously. In the proposed approach, each author in a venue is represented as a probability distribution over topics, and each topic is represented as a probability distribution over words and year of the venue for that topic. Through discovered latent topics we can search mavens by implicitly modeling word-author, author-author and author-venue correlations with continuous time effects. Inference making procedure for topics and authors of new venues is explained. We also show how authors' correlations can be discovered and the bad effect of topics sparseness on the retrieval performance. Experimental results on the corpus downloaded from DBLP show that proposed approach significantly outperformed the baseline approach, due to its ability to produce less sparse topics.

Keywords: Maven Search, Research Community, Topic Modeling, Unsupervised Learning.

1 Introduction

Web is unanimously the biggest source of structured, semi-structured and unstructured data and automatic acquirement of useful information from the web is interesting and challenging from academic recommendation point of view. With the advancement of information retrieval technologies from traditional document-level to object-level [14], expert search 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. Such as, to find appropriate

Q. Li et al. (Eds.): APWeb/WAIM 2009, LNCS 5446, pp. 138–149, 2009.

collaborators for the project, choose mavens for consultation about research topics, reviewers matching for research papers, and to invite program committee members and distinguished speakers for the conferences.

TREC has provided a common platform for researchers to empirically assess approaches for maven search. Several approaches have been proposed to handle this problem. In particular, Cao et al. [6] 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 maven search in those documents. Balog et al. [3] proposed a model which models candidates using its support documents by using language modeling framework. He also proposed several advanced models for maven search specific to sparse data environments [4].

Based on the methods employed in previous language models, they can be classified into two categories: *composite* and *hybrid.* In composite approach, $D_m = \{d_j\}$ denotes the support documents of candidate author r. Each support document d_j is viewed as a unit and the estimation of all the documents of a candidate r are combined. While hybrid approach is very much similar to the composite model, except that it describes each term t_i by using a combination of support documents models and then used a language model to integrate them. Composite approaches suffers from the limitation that all the query terms should occur in one support document and hybrid approaches suffers from the limitation that all the query terms should occur in the support documents.

Generally speaking, language models are lexical-level and ignores semantics-based information present in the documents. Latent semantic structure can be used to capture semantics-based text information for improved knowledge discovery. The idea of using latent semantic structure traces back to Latent Semantic Analysis (LSA) [8], which can map the data using a Singular Value Decomposition (SVD) from high dimensional vector space to reduced lower representation, which is a so-called latent semantic space.

LSA lacks strong statistical foundation; consequently Hofmann [11] proposed probabilistic latent semantic analysis (PLSA). It was based on likelihood principal and defines a proper generative model for the data. By using PLSA, Zhang et al. [20] proposed a mixture model for maven search. They used latent topics between query and support documents to model the mavens by considering all documents of one author as one virtual document. So far all approaches only used authors and supported documents information. While, Tang et al. argued the importance of venues by saying that the documents, authors and venues all are dependent on each other and should be modeled together to obtain the combined influence [17]. Based on it they proposed a unified topic modeling approach by extending Latent Dirichlet Allocation (LDA) [5]. However to the limitation of their approach, they considered the venues information just as a stamp and did not utilize semantics-based text and author's correlations present between the venues.

Previous approaches ignored venues internal semantic structure, author's correlations and time effects. Firstly in real world, venues internal semantic structure and authors correlations are very important, as authors of the renowned venues are more likely to be mavens than authors of not renowned venues. Additionally, renowned venues are more dedicated to specific research areas than not renowned venues, e.g. in famous conferences submission are carefully judged for relevance to the conference

areas, while in not renowned venues it is usually ignored by saying that the topics are not limited to above mentioned research areas on the call for papers page. Some people may think that one has to use impact factors of venues to influence the ranking of mavens, but unluckily there is no standard dataset available to the best of our knowledge. Secondly, continuous-time effects can be very handy in a case if author changes his research interests. For example, an author A was focusing on biological data networks until 2004 and published a lot of papers about this topic; afterwards he switched to academics social network mining and not published many papers. He still can be found as a biological data networks expert in 2008 if we ignore time effects, while it is not an appropriate choice now. The reason for this is occurrence of biological word many times in 2004 and preceding years. However, by attaching time stamp one can minimize the high rate of occurrence effect of one word (e.g. biological) for all years.

In this paper, we investigate the problem of maven search by modeling venues internal semantic structure, author's correlations and time all together. We generalized previous topic modeling approach [17] from a single document to all publications of the venues and added continuous-time considerations, which can provide ranking of mavens in different groups on the basis of semantics. We empirically showed that our approach can clearly produce better results than baseline approach due to topics denseness effect on retrieval performance. We can say that the solution provided by us is well justified and produced quite promising and functional results.

The novelty of work described in this paper lies in the; formalization of the maven search problem, generalization of previous topic modeling approach from document level to venue level (STMS approach) with embedded time effects, and experimental verification of the effectiveness of our approach on the real world corpus. To the best of our knowledge, we are the first to deal with the maven search problem by proposing a generalized topic modeling approach, which can capture word-author, author-author and author-venue correlations with non-discretized time effects.

The rest of the paper is organized as follows. In Section 2, we formalize maven search problem. Section 3 provides maven search modeling related models and illustrates our proposed approach for modeling mavens with its parameter estimation details. In Section 4, corpus, experimental setup, performance measures with empirical studies and discussions about the results are given. Section 5 brings this paper to the conclusions.

2 Maven Search in Research Community

Maven search addresses the task of finding the right person related to a specific knowledge domain. It is becoming one of the biggest challenges for information management in research communities [10]. The question can be like "Who are the mavens on topic *Z*?" A submitted query by user is denoted by *q* and a maven is denoted by *m*. In general semantics-based maven finding process, main task is to probabilistically rank discovered mavens for a given query, i.e. $p\ (m/q)$ where a query is usually comprised of several words or terms.

Our work is focused on finding mavens by using a generalized topic modeling approach. Each conference accepts many papers every year written by different authors.

To our interest, each publication contains some title words and names, which usually covers most of the highly related sub research areas of conferences and authors, respectively. Conferences (or journals) with their accepted papers on the basis of latent topics based correlations can help us to discover mavens. We think that latent topics based correlations between the authors publishing papers in the specific venues by considering time effects is an appropriate way to find mavens. Our thinking is supported by the facts that 1) in highly ranked venues usually papers of mavens or potential mavens of different fields are accepted so venues internal topic-based author correlations are highly influential, 2) highly ranked venues are the best source for analyzing the topical trends due to the reason of mostly accepting and highlighting relatively new ideas, and 3) all accepted papers are very carefully judged for relevance to the venue research areas, so papers are more typical (strongly semantically related).

We denote a venue c as a vector of N_c words based on the paper accepted by the venue for a specific year y, an author r on the basis of his accepted paper (s), and formulize maven search problem as: Given a venue c with N_c words having a stamp of year y, and $\mathbf{a}_c$ authors of a venue c, discover most skilled persons of a specific domain. Formally for finding specific area mavens, we need to calculate the probability $P(z|m)$ and $P(w|z)$ where z is a latent topic, m is maven and w is the words of a venue.

3 Maven Search Modeling

In this section, before describing our STMS approach, we will first describe how mavens can be modeled by using Language Model (LM) [19], Probabilistic Latent Semantic Analysis (PLSA) [11], and Author-Conference-Topic (ACT) Model [17] and why our approach is indispensable.

3.1 Related Models

LM is one of the state-of-the-art modeling approaches for information retrieval. The basic idea is to relate a document given to a query by using the probability of generating a query from given document. In eq. 1, w is a query word token, d is a document and $P(q|d)$ is the probability of the document model generating a query. $P(w|d)$ is the probability of the document model generating a word by using a bag of words assumption.

$$P(q|d) = \prod_{w=q} P(w|d) \tag{1}$$

In eq. 1 one can simply merge all support documents of one author and treat it as a virtual document r representing that author [3,15]. Mavens discovered and ranked with respect to a specific query can be retrieved by using the following equation.

$$P(q|m) = \prod_{w=q} P(w|r) \tag{2}$$

Document and object retrieval with the help of LM has gained a lot of success. But LM faces the inability to exactly match query with the support documents. Hofmann proposed PLSA [11] with a latent layer between query and documents to overcome LM inability by semantically retrieving documents related to a query. The core of

PLSA is a statistical model which is called aspect model [12]. The aspect model is a latent variable model for co-occurrence data which associates an unobserved class variable $z \in Z = \{z_1, z_2, \ldots, z_T\}$ with each observation. A joint probability model over d x w is defined by the mixture, where, each pair (d,w) is assumed to be generated independently, corresponding to bag of words assumption words w are generated independently for the specific document d conditioned on topic z.

$$P(d,w) = P(d)P(w|d),$$
$$where\ P(w|d) = \sum_{z=Z} P(w|z)P(z|d) \tag{3}$$

In eq. 3 by changing word w with query q and document d with author r, where r can be seen as a composition of all documents of one author as one virtual document [20]. Here, P (m) can be calculated by a variety of techniques [4,21] to obtain the joint probability of maven and query which can be used to discover and rank mavens.

$$P(m,q) = P(r)P(q|r),$$
$$where\ P(q|r) = \sum_{z=Z} P(q|z)P(z|r) \tag{4}$$

Recently, Author-Conference-Topic (ACT) model was proposed for expertise search [17]. In ACT model, each author is represented by the probability distribution over topics and each topic is represented as a probability distribution over words and venues for each word of a document for that topic. Here venue is viewed as a stamp associated with each word with same value. Therefore, the modeling is just based on semantics-based text information and co-authorship of documents, while semantics-based intrinsic structure of words and authors' correlations with embedding time effects present in the venues on the basis of writing for the same venue are ignored, which became the reason of topic sparseness that resulted in poor retrieval performance. The generative probability of the word w with venue c for author r of a document d is given as:

$$P(w,c|r,d,\emptyset,\Psi,\theta) = \sum_{z=1}^{T} P(w|z,\emptyset_z)P(c|z,\Psi_z)P(z|r,\theta_r) \tag{5}$$

3.2 Semantics and Temporal Information Based Maven Search (STMS) Approach

We think it is necessary to model venues internal semantic structure and author correlations than only considering venues as a stamp [17] and time factor for maven search. The basic idea presented in Author-Topic model [16], that words and authors of the documents can be modeled by considering latent topics became the intuition of modeling words, authors, venues and years, simultaneously. We generalized the idea presented in [17] from documents level (DL) to venues level (VL) by considering research papers as sub-entities of the venues to model the influence of renowned and not renowned venues on the basis of participation in same venues. Additionally, we considered continuous-time factor to deal with the topic drift in different years. In the proposed approach, we viewed a venue as a composition of documents words and the authors of its accepted publications with year as a stamp. Symbolically, for a venue c (a virtual document) we can write it as: $C = [\{(\mathbf{d}_1,\mathbf{a}_{d1}) + (\mathbf{d}_2,\mathbf{a}_{d2})+(\mathbf{d}_3,\mathbf{a}_{d3})+ \ldots + (\mathbf{d}_i,\mathbf{a}_{di})\} + y_c]$ where $\mathbf{d}_i$ is a word vector of document published in a venue, $\mathbf{a}_{di}$ is author vector of d_i and y_c is paper publishing year.

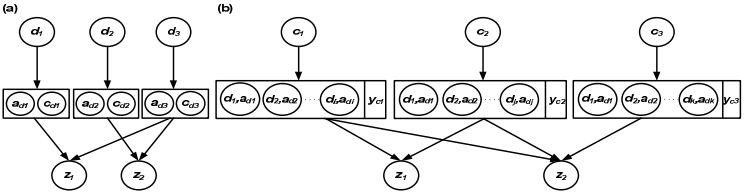

Fig. 1. Semantics-Based Maven search a) ACT and b) STMS approaches

DL approach considers that an author is responsible for generating some latent topics of the documents on the basis of semantics-based text information and co-authorship. While, VL approach considers that an author is responsible for generating some latent topics of the venues on the basis of semantics-based text information and authors correlations with time is not-discretized (please see fig. 1). In STMS approach, each author (from set of K authors) of a venue 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 a venue 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 venue c is given as:

$$P(w, y|r, c, \emptyset, \Psi, \theta) = \sum_{z=1}^{T} P(w|z, \emptyset_z) P(y|z, \Psi_z) P(z|r, \theta_r) \quad (6)$$

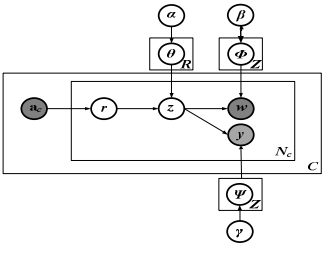

Fig. 2. STMS approach

The generative process of STMS is as follows:

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

Gibbs sampling is utilized [1,9] for parameter estimation in our approach, which has two latent variables z and r; the conditional posterior distribution for z and r is given by:

$$P(z_i = j,\ r_i = k | w_i = m, y_i = n,\ \mathbf{z}_{-i}, \boldsymbol{r}_{-i}, \mathbf{a}_c) \propto \frac{n_{-i,j}^{(wi)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(yi)} + \gamma}{n_{-i,j}^{(\cdot)} + Y\gamma} \frac{n_{-i,j}^{(ri)} + \alpha}{n_{-i,\cdot}^{(ri)} + R\alpha} \tag{7}$$

where $z_i = j$ and $r_i = k$ represent the assignments of the i^{th} word in a venue to a topic j and author k respectively, $w_i = m$ represents the observation that i^{th} word is the m^{th} word in the lexicon, $y_i = n$ represents i^{th} 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 i^{th} 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. "." 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 x Z (word by topic), Y x Z (year by topic) and Z x R (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:

$$\emptyset_{zw} = \frac{n_{-i,j}^{(wi)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \tag{8}$$

$$\Psi_{zy} = \frac{n_{-i,j}^{(yi)} + \gamma}{n_{-i,j}^{(\cdot)} + Y\gamma} \tag{9}$$

$$\theta_{rz} = \frac{n_{-i,j}^{(ri)} + \alpha}{n_{-i,\cdot}^{(ri)} + R\alpha} \tag{10}$$

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. The mavens related to a query can be found and ranked with respect to their probabilities as:

$$P(m|q) \propto P(q|m) = \prod_{w \in q} \sum_{z \in Z} P(w|z) P(z|r) \tag{11}$$

Where, w is words contained in a query q and m denotes a maven.

4 Experiments

4.1 Corpus

We downloaded five years publication Corpus of venues from DBLP [7], by only considering conferences for which data was available for years 2003-2007. In total,

we extracted 112,317 authors, 62,563 publications, and combined them into a virtual document separately for 261 conferences each year. 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.

4.2 Parameter Setting

In our experiments, for 150 topics Z the hyper-parameters α, β and γ were set at 50/Z, .01, and 0.1. Topics are set at 150 at a minimized perplexity [5], a standard measure for estimating the performance of probabilistic models with the lower the best, for the estimated topic models. Teh et al. proposed a solution for automatic selection of number of topics, which can also be used for topic optimization [18].

4.3 Performance Measures

Perplexity is usually used to measure the performance of latent-topic based approaches; however it cannot be a statistically significant measure when they are used for information retrieval [Please see [2] for details]. In our experiments, firstly we use average entropy to measure the quality of discovered topics, which reveals the purity of topics. Entropy is a measure of the disorder of system, less intra-topic entropy is better. Secondly, we used average Symmetric KL (sKL) divergence [17,20] to measure the quality of topics, in terms of inter-topic distance. sKL divergence is used here to measure the relationship between two topics, more inter-topic sKL divergence (distance) is better.

$$Entropy\ of\ (Topic) = -\sum_z P(z) log_2[P(z)] \tag{12}$$

$$sKL(i,j) = \sum_{z=1}^{T} \left[\theta_{iz} log \frac{\theta_{iz}}{\theta_{jz}} + \theta_{jz} log \frac{\theta_{jz}}{\theta_{iz}} \right] \tag{13}$$

4.4 Baseline Approach

We compared proposed STMS approach with ACT approach and used same number of topics for comparability. The number of Gibbs sampler iterations used for ACT is 1000 and parameter values same as the values used in [17].

4.5 Results and Discussions

4.5.1 Topically Related Mavens

We extracted and probabilistically ranked mavens related to a specific area of research on the basis of latent topics. Tab. 1 illustrates 5 different topics out of 150, discovered from the 100th iteration of the particular Gibbs sampler run.

Table 1. An illustration of 5 topics with related mavens. The titles are our interpretation of the topics.

Topic 27 "XML Databases"		Topic 123 "Software Engineering"		Topic 98 "Robotics"		Topic 119 "Data Mining"		Topic 35 "Bayesian Learning"	
Word	***Prob.***	***Word***	***Word***	***Word***	***Prob.***	***Word***	***Prob.***	***Word***	***Prob.***
Data	0.110127	Software	0.083778	Robot	0.093050	Mining	0.164206	Learning	0.146662
XML	0.068910	Development	0.035603	Control	0.041873	Data	0.10328	Bayesian	0.039022
Query	0.047482	Oriented	0.031238	Robots	0.039397	Clustering	0.064461	Models	0.030610
Databases	0.038257	Engineering	0.028553	Motion	0.032675	Patterns	0.039122	Classification	0.027550
Database	0.037364	Systems	0.022006	Robotic	0.021119	Frequent	0.030041	Markov	0.018373
Queries	0.034239	Model	0.018649	Planning	0.018761	Series	0.023889	Kernel	0.015123
Processing	0.027543	Component	0.017306	Force	0.016638	Streams	0.023010	Semi	0.013785
Relational	0.020401	Tool	0.016467	Tracking	0.015695	Dimensional	0.012611	Regression	0.013785
Mavens	***Prob.***	***Mavens***	***Author***	***Mavens***	***Prob.***	***Mavens***	***Prob.***	***Mavens***	***Prob.***
Divesh Srivastava	0.011326	Gerardo Canfora	0.012056	Gerd Hirzinger	0.014581	Philip S. Yu	0.021991	Zoubin Ghahramani	0.012506
Elke A. Rundensteiner	0.010205	Tao Xie	0.009842	Paolo Dario	0.013586	Wei Wang	0.01785	Andrew Y. Ng	0.011882
Kian-Lee Tan	0.009747	Jun Han	0.005930	Ning Xi	0.011657	Jiawei Han	0.017792	John Langford	0.006266
Tok Wang Ling	0.008881	Nenad Medvidovic	0.005192	Toshio Fukuda	0.011470	Hans-Peter Kriegel	0.014686	Michael H. Bowling	0.006032
Surajit Chaudhuri	0.008779	Johannes Mayer	0.004970	Atsuo Takanishi	0.010475	Christos Faloutsos	0.01066	Sanjay Jain	0.005798
Rakesh Agrawal	0.008269	Jason O. Hallstrom	0.004601	Vijay Kumar	0.008857	Reda Alhajj	0.010373	Harry Zhang	0.005486
Sharma Chakravarthy	0.008218	S. C. Cheung	0.004158	Yoshihiko Naka.	0.007737	Eamonn J. Keogh	0.009567	Doina Precup	0.005018
Jeffrey F. Naughton	0.007913	Lu Zhang	0.003715	Joel W. Burdick	0.007426	Jian Pei	0.008992	Kai Yu	0.004706

The words associated with each topic are quite intuitive and precise in the sense of conveying a semantic summary of a specific area of research. For example, topic # 27 "XML Databases" shows quite specific and precise words when a person is searching for databases experts with move from simple databases to XML databases. Other topics shown in the tab. 1 are also quite descriptive that shows the ability of STMS approach to discover compact topics. The mavens associated with each topic are quite representative, as we have analyzed that, top ranked mavens for different topics are typically mavens of that area of research. For example, in case of topic 35 "Bayesian Learning" and topic 119 "Data Mining" top ranked mavens are well-known in their respective fields.

Proposed approach discovered several other topics related to data mining such as neural networks, multi-agent systems and pattern recognition, also other topics that span the full range of areas encompassed in the corpus.

In addition, by doing analysis of mavens home pages and DBLP [7], we found that 1) all highly ranked mavens have evenly published papers on their relevant topics for all years, no matter they are old or new researchers and 2) all of their papers are usually published in the well-known venues. Both findings provide qualitative supporting evidence for the effectiveness of the proposed approach.

Fig. 3 provides a quantitative comparison between STMS and ACT models. Fig. 3 (a) shows the average entropy of topic-word distribution for all topics calculated by

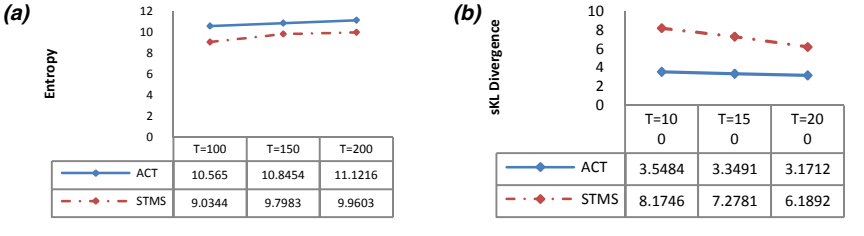

Fig. 3. a) Average Entropy curve as a function of different number of topics and **b)** Average sKL divergence curve as a function of different number of topics

using eq. 12. Lower entropy for different number of topics T= 100, 150, 200 proves the effectiveness of proposed approach for obtaining dense (less sparse, clearer) topics. Fig. 3 (b) shows the average distance of topic-word distribution between all pairs of the topics calculated by using eq. 13. Higher sKL divergence for different number of topics T= 100, 150, 200 confirms the effectiveness of proposed approach for obtaining dense topics.

One would like to quickly acquire the topics and mavens for new venues that are not contained in the training corpus. Provided parameter estimation Gibbs sampling algorithm requires significant processing time for large number of dataset. It is computationally inefficient to rerun the Gibbs sampling algorithm for every new venue added to the corpus. For this purpose, we can apply eq. 7 only on the word tokens and authors of the new venue each time temporarily updating the count matrices of (word by topic) and (topic by author). The resulting assignments of words to topics can be saved after a few iterations (10 in our simulations) and then eq. 11 can be used to search query related mavens.

4.5.2 Effect of Topic Sparseness on Retrieval Performance

Topic by author matrix (influenced by venues and time information) can also be used for automatic correlation discovery between authors more appropriately, than previously used topic by author matrix (not influenced by venues and time information) [16]. To illustrate how it can be used in this respect, distance between authors *i* and *j* is calculated by using eq. 13 for topic-author distribution.

We calculated the dissimilarity between the authors; smaller dissimilarity value means higher correlation between the authors. Tab. 2 shows top 7 semantics-based mavens ids discovered related to the first maven of each topic for STMS and ACT approaches. For example, in case of "XML Databases" topic 4808, 337, 5194, 4457, 4870, 4775, 640 are top 7 mavens correlated with "Divesh Srivastava" for STMS approach in terms of sKL divergence and so on.

Table 2. An illustration of 5 topics sparseness effect on retrieval performance in terms of error rate (ER)

STMS Approach					ACT Approach				
XML Databases	**Software Engineering**	**Robotics**	**Data Mining**	**Bayesian Learning**	**XML Databases**	**Software Engineering**	**Robotics**	**Data Mining**	**Bayesian Learning**
4808	6871	12723	4477	11094	9398	12823	24645	14131	3627
337	14588	12508	5119	3289	14221	6700	24952	14409	19973
5194	2531	12887	2644	924	14401	3403	24828	1467	3655
4457	19610	1898	4743	3250	13696	7786	24808	1499	19988
4870	832	12496	10282	1877	6275	7637	24699	14589	23912
4775	13304	12486	10326	9637	13620	2525	9202	4410	5922
640	25680	4915	323	1682	14248	18352	24643	815	10974
ER=57.14	ER=57.14	ER=71.43	ER=42.85	ER=28.57	ER=71.43	ER=71.43	ER=71.43	ER=71.43	ER=57.14
Average Error Rate = 51.43					**Average Error Rate = 68.57**				

The highlighted blocks in tab. 2 shows that similar results are found for discovered topics and sKL divergence. For example, in case of STMS approach top eight mavens shown in tab. 1 for "XML Databases" topic has three mavens in common, which are 337 "Sharma Chakravarthy", 4457 "Tok Wang Ling", and 4775 "Surajit Chaudhuri". From top 7 related mavens for five selected topics (same is the case with non selected topics) shown in the tab. 1 the error rate (ER) for STMS is less than ACT and STMS

has 17.14 % less average error rate than ACT. It shows the bad effect of topics sparseness on maven retrieval performance for ACT, and inability of ACT to discover better results in comparison with STMS.

5 Conclusions and Future Works

This study deals with the problem of maven search through latent topics. Initially we generalized this problem to VL with embedding continuous time effects and discussed the motivation for it. We then introduced STMS approach, which can discover and probabilistically rank experts related to specific knowledge domains (or queries) by modeling semantics-based correlations and temporal information simultaneously. We demonstrated how it can be used to rank experts for unseen data and to find mavens correlations. We studied the effect of generalization on topics denseness when modeling entities and concluded that more dense topics will results in better performance of the approach. Empirical results show better performance on the basis of compact topics as compared to the baseline approach. As a future work, we plan to investigate how to use STMS approach for ranking mavens related to a topic for different years and discover changing trends in their expertness, as we think for different years the mavens are usually different and the expertness of an author can be dynamic over different time span.

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

References

1. Andrieu, C., Freitas, N.D., Doucet, A., Jordan, M.: An Introduction to MCMC for Machine Learning. Journal of Machine Learning 50, 5–43 (2003)
2. Azzopardi, L., Girolami, M., Risjbergen, K.V.: Investigating the Relationship between Language Model Perplexity and IR Precision-Recall Measures. In: Proc. of the 26th ACM SIGIR, Toronto, Canada, July 28-August 1 (2003)
3. Balog, K., Azzopardi, L., de Rijke, M.: Formal Models for Expert Finding in Enterprise Corpora. In: Proc. of SIGIR, pp. 43–55 (2006)
4. Balog, K., Bogers, T., Azzopardi, L., Rijke, M., Bosch, A.: Broad Expertise Retrieval in Sparse Data Environments. In: Proc. of SIGIR, pp. 551–558 (2007)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
6. Cao, Y., Liu, J., Bao, S., Li, H.: Research on Expert Search at Enterprise Track of TREC (2005)
7. DBLP Bibliography Database, `http://www.informatik.uni-trier.de/~ley/db/`
8. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)

9. Griffiths, T.L., Steyvers, M.: Finding Scientific Topics. In: Proc. of the National Academy of Sciences, USA, pp. 5228–5235 (2004)
10. Hawking, D.: Challenges in Enterprise Search. In: Proc. of the 15th Conference on Australasian Database, vol. 27, pp. 15–24 (2004)
11. Hofmann, T.: Probabilistic Latent Semantic Analysis. In: Proc. of the 15th Annual Conference on UAI, Stockholm, Sweden, July 30-August 1 (1999)
12. Hofmann, T., Puzicha, J., Jordan, M.I.: Learning from Dyadic Data. In: Advances in Neural Information Processing Systems (NIPS), vol. 11. MIT Press, Cambridge (1999)
13. Mimno, D., McCallum, A.: Expertise Modeling for Matching Papers with Reviewers. In: Proc. of the 13th ACM SIGKDD, pp. 500–509 (2007)
14. Nie, Z., Ma, Y., Shi, S., Wen, J., Ma, W.: Web Object Retrieval. In: Proc. of World Wide Web (WWW), pp. 81–90 (2007)
15. Petkova, D., Croft, W.B.: Generalizing the Language Modeling Framework for Named Entity Retrieval. In: Proc. of SIGIR (2007)
16. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The Author-Topic Model for Authors and Documents. In: Proc. of the 20th International Conference on UAI, Canada (2004)
17. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: Extraction and Mining of Academic Social Networks. In: Proc. of the 14th ACM SIGKDD (2008)
18. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet Processes. Technical Report 653, Department of Statistics, UC Berkeley (2004)
19. Zhai, C., Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Ad-hoc Information Retrieval. In: Proc. of the 24th ACM SIGIR, pp. 334–342 (2001)
20. Zhang, J., Tang, J., Liu, L., Li, J.: A Mixture Model for Expert Finding. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 466–478. Springer, Heidelberg (2008)
21. Zhang, J., Tang, J., Li, J.: Expert Finding in a Social Network. In: Kotagiri, R., Radha Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 1066–1069. Springer, Heidelberg (2007)