Data & Knowledge Engineering 69 (2010) 965–978

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Modeling the evolution of associated data

Jie Tang *, Jing Zhang

Department of Computer Science and Technology, Tsinghua University, China

ARTICLE INFO

Article history:
Received 15 September 2008
Received in revised form 16 March 2010
Accepted 16 March 2010
Available online 25 March 2010

Keywords:
Topic model
Probabilistic model
Evolution analysis
Knowledge discovery

ABSTRACT

Statistical topic models have been proposed for modeling documents and authorship information. However, few previous works have studied the evolution of associated data. In this paper, we investigate how to model trends of changes in document content and author interests simultaneously over time. We propose two models: a bag-of-words based Author–Time–Topic model that extends the state-of-the-art LDA-style topic model and a Hidden Markov Author–Time–Topic model, which can model interdependencies between topics. We use the Gibbs EM algorithm for parameter estimation. We apply these models to two data sets: NIPS papers and Yahoo group posts. Experimental results show that our models can achieve a lower perplexity (−2.0%–20%) than the baseline LDA and Author–Topic model, when modeling quickly evolving associated data. Experiments also reveal that the proposed models can accurately capture the hot topics in different periods (e.g. "Yao at preseason" in Aug-2004, when the Chinese player Ming Yao became a highlight in the NBA) from the two data sets.

1. Introduction

Characterizing different types of information (e.g. document content and authorship) contained in rapidly growing electronic document collections can benefit many data mining applications. For example, finding topics from document contents is a standard problem in information retrieval, natural language processing, and machine learning; while modeling the interests of authors can be utilized to answer a range of important questions, such as which subjects an author is interested in and who are likely to be the experts on a given subject.

From another point of view, the requirements for modeling trends of changes in associated data (e.g., documents and authors' interests) are also becoming more and more important. For example, changes of the themes of documents that a researcher is working on would reflect the shifts of his interest.

A recommendation system can use the evolution analysis results of the associated data to reinforce the recommendation results. Many large data sets on the current Web are dynamic rather than static. Moreover, the change patterns of different types of information are usually interdependent. Consequently, it becomes a challenge to model the evolution trends of different forms of information simultaneously in a unified model. This is the problem addressed in this paper.

Many topic models have been proposed for modeling document contents and author interests, for example, LDA (a topic model) [5] [15], the Author model [19], and the Author–Topic model [31] [32]. Latent Dirichlet Allocation (LDA) has been proposed to discover multinomial word distribution over topics. The Author model is similar to LDA. The difference is that it is used for modeling a mixture of word distributions over authors. The Author–Topic model is then proposed to model document contents and authors' interests. However, all of the aforementioned models ignore important factor–time information, which reveals a huge amount of

* Corresponding author. Rm 1-308, FIT Building, Tsinghua University, Beijing, 100084, China. Tel.: +86 10 62788788 20; fax: +86 10 62794365.
E-mail addresses: jietang@tsinghua.edu.cn (J. Tang), zhangjing@keg.cs.tsinghua.edu.cn (J. Zhang).

0169-023X/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2010.03.009

information contained in document collections. Many real-world applications, for example topic detection and tracking (TDT), research trend analysis on scientific papers, and hot topic finding from newsgroup posts need to consider the evolution of topics over time.

Recently, several efforts have been made at integrating time information into topic models. For example, topics over time (TOT) [40] associates with each topic a Beta distribution that represents the occurrence probability of the topic at a given time. The Dynamic Topic Models (DTM) [4] model the evolution of topics by estimating the topic distribution at various epochs. However, no previous work has simultaneously modeled change trends of associated data such as documents and authors' interests.

Several questions arise in modeling associated data: (1) is there a way to model the change trends of different types of data simultaneously? (2) Is there a better way to model the textual information beyond the unigram model? (3) Can information on the evolution of different data be combined in useful ways? To the best of our knowledge, these problems have not yet been seriously investigated.

In this paper, we present two generative models, i.e., Author–Time–Topic (ATT) model and Hidden Markov Author–Time–Topic (HMATT) model, for simultaneously modeling trends of changes in document contents and authors' interests. The ATT model represents each document by a mixture of topics and assigns a mixture of weights of different topics to the authors of a document. To model topic distributions, we associate to each topic a continuous distribution over time, so that topics are responsible for generating both observed words and timestamps. The other model, i.e., Hidden Markov Author–Time–Topic, is proposed by further considering dependencies between topics. In HMATT, the topics of words in a document are viewed as a Markov chain and thus the model captures interdependencies between topics. We present experimental results on two real-world data sets: NIPS papers and Yahoo Group posts. Experimental results show that clear improvements in the trend analysis of author interests and document characteristics can be obtained by the ATT and HMATT models, compared with the LDA and AT models.

The contributions of this paper include: (a) the proposal of the ATT model for simultaneously modeling trends of changes in document contents and author interests, (b) the proposed application of the HMATT model to the problem by further considering dependencies between topics and (c) the empirical verification of the effectiveness of the proposed models.

The rest of the paper is organized as follows. In Section 2, we introduce related work from the literature. In Section 3 we present the proposed ATT and HMATT models. In Section 4, we explain the algorithms for inference and in Section 5, we discuss the experimental results. Section 6 concludes this work.

2. Related work

Many models have been proposed for modeling documents, for example, language models, probabilistic Latent Semantic Indexing models (pLSI) [15] and Latent Dirichlet Allocation (LDA) [5]. In this section, we review three aspects of the related work: topic evolution modeling, author interest modeling, and topic dependency modeling.

2.1. Modeling topic evolution

Several researchers have studied topic evolution in the recent past. The previous models either predivide the data into discrete time slices or consider continuous time without discretization.

For instance, for time-discretization based modeling, Blei and Lafferty propose Dynamic Topic Models (DTMs) in which the alignment among topics across time slices is captured by a Kalman filter [4]. However, as the model uses a normal distribution that is not a conjugate to the multinomial distribution, the model does not yield a simple solution to the inference. Nallapati et al. propose a Multiscale Topic Tomography Model (MTTM) [26], which employs inhomogeneous Poisson processes to model the generation of word-counts. The evolution of topics is modeled through a multiscale analysis using Haar wavelets. Quon et al. [30] study the problem of evolution of gene expression in inhomogeneous data sets, and present a statistical model to characterize changes in expression among highly complex organisms.

Another category of models directly modeling the continuous time information. For example, Nodelman et al. propose Continuous Time Bayesian Networks (CTBN) to model continuous time flow based on the Markov assumption without discretization [28]. Zhao et al. [41] also propose a method for event detection from evolution of click-through data.

However, all of these models only consider modeling the evolution of one type of information. To the best of our knowledge, no previous work has simultaneously modeled trends of change in different types of information.

2.2. Modeling author interest with topics

Some other efforts have been made at modeling authors' interests. For example, the Author model (also called the Multilabel Mixture Model) [19] is intended to model author interests with a one-to-one correspondence between topics and authors. In [31] [32], an Author–Topic model is presented, which integrates authorship into the topic model and thus can be used to find a topic distribution over document and a mixture of the distributions associated with authors.

The Author–Topic (AT) model is a Bayesian network similar to that in LDA [5]. In the AT model, each author's interests is modeled with a mixture of topics [31] [32]. To generate a word of a document, an author is first chosen uniformly at random from the authors, then a topic is selected from a topic distribution specific to the author, and then a word is generated by sampling from the chosen topic. McCallum et al. have studied several other topic models in social network analysis [20]. They propose the

Author–Recipient–Topic (ART) model, which learns topic distributions based on emails sent between people. The topic distribution is conditioned on the email contents, senders, and recipients. Leskovec et al. [17] propose a method for tracking new topics, ideas, and "memes" across the Web.

In recent years, a few works have been conducted for modeling the linked data. For example, Ahmed et al. [1] propose a structured correspondence topic model for mining figure captions in the biological literature. Nallapati et al. [25] present two different topic models for joint modeling of text and citations. Liu et al. [18] propose a topic modeling approach to predict the link relationship between documents. Gruber et al. [14] propose a latent topic model for Hypertext. There are considerable efforts that have been made to extend the topic model. Doyle and Elkan [9] extend the topic model to discover burstiness; Iwata et al. [16] propose using a topic modeling approach to model the social annotation data.

However, these models cannot model time information, and thus cannot model trends of changes in information.

2.3. Modeling dependencies between topics

There are also a few other models related to the proposed HMATT model. For example, Wallach proposes a Bigram Topic model [38], which incorporates both n-gram statistics and latent topics by extending a unigram topic model to include properties of a hierarchical Dirichlet bigram model. Tang et al. [35] present a method to discover topic distribution over links between pages. Deschacht and Moens [7] present a latent words language model, which dependencies between latent words are taken into account for semi-supervised semantic role labeling.

In [12], a model that integrates topics and syntax is introduced. It contains a latent variable for each word that stands for syntactic classes. However, the model assumes that each word is generated either from a latent topic or from a syntactic class. Only the syntactic classes are treated as a sequence with local dependencies while the latent topics are not. Hidden Topic Markov Model [13] is based on a strong assumption that all words in the same sentence have the same topic. Putthividhya et al. [29] improve upon the correlated topic model (CTM) [3] and propose Independent Factor Topic Models (IFTM) which use linear latent variable models to uncover the hidden sources of correlation between topics. Boyd-Graber and Blei [6] further study a nonparametric Bayesian model for modeling documents by considering syntactical information available from parsing trees generated by a natural language parser.

The probabilistic topic models have been widely applied to various applications such as ontology learning [33] [39] and clustering distributed databases [21]. Several recent studies also consider how to model the imbalanced and noisy data [36] and how to capture the interactions of subsets of feature values for instance classifications [37].

3. The proposed topic models

Table 1 summarizes the notations used throughout this paper. Given a collection of documents $\boldsymbol{D}=\{(\boldsymbol{w}_1, \boldsymbol{a}_1, t_1), \cdots, (\boldsymbol{w}_D, \boldsymbol{a}_D, t_D)\}$, where $\boldsymbol{w}_d$ denotes the sequence of N_d words in document d, $\boldsymbol{a}_d$ denotes a vector of A_d authors of document d, and t_d denotes the timestamp of document d. Each $w_{di} \in \boldsymbol{w}_d$ is chosen from a vocabulary of size V and each author $a_{di} \in \boldsymbol{a}_d$ is chosen from a set of authors of size A. In addition, let x_{di} indicate an author, chosen from $\boldsymbol{a}_d$, responsible for the ith word in document d. Here each author is associated with a distribution over topics θ, chosen from a symmetric Dirichlet(α) prior distribution. The number of topics is denoted by T.

Modeling the evolution of multiple related objects such as documents' contents and authors' interests is a critical issue in many applications. Traditionally, different objects are modeled separately, and documents are usually modeled based on the "bag-of-words" (BOW) assumption. However, such an approach cannot take advantage the "semantic" dependencies between different types of objects and cannot capture the relationships between words.

Table 1
Notations.

Symbol	Description
T	Number of topics
D	Number of documents
V	Number of unique words
A	Number of unique authors
M	Number of timestamps
N_d	Number of word tokens in document d
$\boldsymbol{w}_d$	Vector form of document d
$\boldsymbol{a}_d$	Vector form of authors in document d
w_{di}	The ith word token in document d
t_{di}	The timestamp associated with w_{di}
θ_d	Multinomial distribution over topics
z_{di}	Topic assigned with word token w_{di}
ϕ_z	Multinomial distribution of words specific to z
ψ_z	Beta distribution of time specific to topic z
x_{di}	The author a associated with w_{di}

To deal with this problem, we propose two unified topic models for simultaneously modeling trends of changes in documents' contents and authors' interests, called Author–Time–Topic (ATT) model and Hidden Markov Author–Time–Topic (HMATT) model. The ATT model combines two generative processes to discover the topic-based change patterns of documents and authors. The HMATT further considers the dependencies within the sequence of words. This is very useful, in particular to model trends of changes in the associated data. For example, the word "learner" may have two topics: one for machine learning and the other for human learners. If the previous word is "decision", "transductive", or "lazy", then the word "learner" would has a high probability to be assigned to be a machine learning topic, whereas if the previous word is "education" or "teach", then the word is more likely to be assigned with a human learning topic.

3.1. Author–Time–Topic model

The Author–Time–Topic (ATT) model combines the AT model [31] and the TOT model [40] to estimate topic distributions simultaneously over words and timestamps. The basic idea of the ATT model is that each word token and its associated timestamp are generated from the same sampled topic, and thus the posterior probability distribution of the sampled topic depends on both the word and the timestamp. The corresponding generative process in the ATT model can be described as

- For each topic z, draw ϕ_z from a Dirichlet prior β_z;
- For each word w_{di} in document d
  - Draw an author x_{di} from $\boldsymbol{a}_d$ uniformly;
  - Draw a topic z_{di} from a multinomial distribution $\theta_{x_{di}}$ specific to author x_{di}, where θ is generated from a Dirichlet prior α;
  - Draw a word w_{di} from a multinomial distribution $\phi_{z_{di}}$;
  - Draw a timestamp t_{di} from a Beta distribution $\text{Beta}(\psi_{z_{di}})$.

The graphical model of ATT is shown in Fig. 1. All timestamps associated to words in a document are observed as the same as the timestamp of the document (e.g., the publication year of a paper).

In the ATT model, the joint probability of words $\boldsymbol{w}$, timestamps $\boldsymbol{t}$, a set of corresponding latent topics $\boldsymbol{z}$, and an author mixture $\boldsymbol{x}$ is defined as

$$P(\mathbf{x},\mathbf{z},\mathbf{w},\mathbf{t}|\Theta,\Phi,\Psi,\mathbf{a}) = \prod_{d=1}^{D}\prod_{i=1}^{N_d}\frac{1}{A_d}\text{Beta}\left(\psi_{z_{di}}^{t_{di}}\right) \times \prod_{z=1}^{T}\prod_{v=1}^{V}\prod_{x=1}^{A}\theta_{xz}^{m_{xz}}\phi_{zv}^{n_{zv}} \tag{1}$$

where m_{xz} is the number of times that topic z has been associated with the chosen author x and n_{zv} is the number of times that word w_v has been generated by topic z. Beta(.) is defined as

$$\text{Beta}\left(\psi_z^{t_{di}}\right) = \frac{(1-t_{di})^{\psi_{z_{di}1}-1}t_{di}^{\psi_{z_{di}2}-1}}{B\left(\psi_{z_{di}1},\psi_{z_{di}2}\right)} \tag{2}$$

where $B(.)$ is a Beta function. We also tried different other distributions for the time information, for example, the same multinomial distribution as for words, which results in a model similar to the Author–Conference–Topic (ACT) model [34]. We found that the multinomial and other distributions cannot produce satisfactory results.

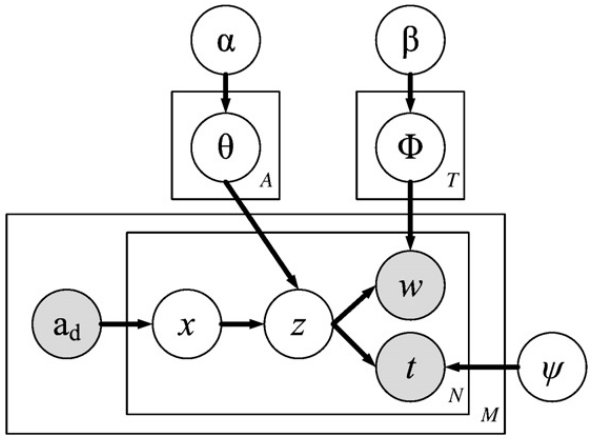


Fig. 1. Graphical model in ATT. $\boldsymbol{a}_d$ is a vector of authors of a document d; x is an author selected to be responsible for a word w, t is the timestamp, and z is the topic assigned to each word. α, β, and ψ are the hyperparameters.

By placing a Dirichlet prior α over Θ and another prior β over Φ, and combining them into Eq. (1), and then further integrating over Θ and Φ, we obtain

$$\begin{aligned} P(\mathbf{x},\mathbf{z},\mathbf{w},\mathbf{t}|\alpha,\beta,\Psi,\mathbf{a}) = & \prod_{d=1}^{D}\prod_{i=1}^{N_d}\frac{1}{A_d}\mathrm{Beta}\left(\psi_{z_{di}}^{t_{di}}\right) \\ & \times \prod_{x=1}^{A}\frac{\Gamma(\sum_z \alpha_z)}{\prod_{z=1}^{T}\Gamma(\alpha_z)}\frac{\prod_{z=1}^{T}\Gamma(m_{xz}+\alpha_z)}{\Gamma(\sum_{z=1}^{T}(m_{xz}+\alpha_z))} \\ & \times \prod_{z=1}^{T}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma(n_{zv}+\beta_v)}{\Gamma(\sum_{v=1}^{V}(n_{zv}+\beta_v))}. \end{aligned} \tag{3}$$

There is a set of unknown parameters in the ATT model: (1) the distribution θ of D document-topics and the distribution ϕ of T topic-words; (2) the distribution ψ of $T\times M$ topic-time and the corresponding topic z_{di} for each word w_{di} in the document d. It is usually intractable to exactly estimate the parameters in such a probabilistic model. A variety of algorithms have been proposed to estimate them approximately, for example variational EM methods [5], Gibbs sampling [11] [32], and expectation propagation [11] [23]. We choose Gibbs sampling for its ease of implementation. Specifically, Gibbs sampling can get around the intractability of directly estimating the model parameters, as instead it first calculates the posterior distribution on just z and then use the results to infer θ, ϕ, and ψ.

Given D documents, a set of topics z, and hyperparameters α and β, the random variables ϕ (the probability of a given word given a topic) and θ (the probability of a given topic given an author) can be estimated via

$$\phi_{zw_{di}} = \frac{n_{zw_{di}}+\beta_{w_{di}}}{\sum_{v=1}^{V}(n_{zv}+\beta_v)} \tag{4}$$

$$\theta_{xz} = \frac{m_{xz}+\alpha_z}{\sum_{z'=1}^{T}(m_{xz'}+\alpha_{z'})}. \tag{5}$$

The random variables ψ_z can be updated after Gibbs sampling by fixing the sampled topics for the words [40]

$$\psi_{z1} = t_z^y\left(\frac{t_z^y(1-t_z^y)}{s_z^2}-1\right) \tag{6}$$

$$\psi_{z2} = (1-t_z^y)\left(\frac{t_z^y(1-t_z^y)}{s_z^2}-1\right) \tag{7}$$

where t_z^y and s_z^2 respectively denote the sample mean and the biased sample variance of the timestamps to topic z.

3.2. Hidden Markov Author–Time–Topic model

The HMATT model further extends the ATT model by considering the dependencies between topics. Traditional LDA-style topic models are usually based on the bag-of-words assumption, and thus ignore dependencies between topics. These models make sense from the perspective of computational efficiency, but are unrealistic. In many language modeling applications, such as named entity recognition and part-of-speech (POS) tagging, dependencies (even strong dependencies) exist between topics. We also found this problem in our setting. This leads us to think about how to incorporate dependencies between topics into the ATT model.

There are several ways of describing the dependencies. We propose a Hidden Markov Author–Time–Topic (HMATT) model, which defines a conditional distribution $P(z_k|z_j x_k)$, described by $(AT\times T)$ free parameters. Their parameters form an AT rows matrix Θ, with $\mathrm{P}(z_k|z_j x_k)=\theta_{(zk|zjxk)}$ Each row is a distribution over topics for a particular context z_j and x_k denoted by Θ_{zjxk}.

Based on this consideration, the generative process for a corpus can be defined as

- For each topic z, draw ϕ_z from a Dirichlet prior β_z;
- For each word w_{di} in document d
  - Draw an author x_{di} from a_d uniformly;
  - Draw a topic z_{di} from a multinomial $\theta_{x_{di}z_{d(i-1)}}$ that is defined by the author x_{di} and previously sampled topic $z_{d(i-1)}$, where θ is generated from a Dirichlet prior α;
  - Draw a word w_{di} from a multinomial distribution $\phi_{z_{di}}$;
  - Draw a timestamp t_{di} from a Beta distribution $\mathrm{Beta}(\psi_{z_{di}})$.

A graphical model of HMATT is shown in Fig. 2. We see that in HMATT, topic generation is determined not only by the chosen author but also by the previous topic. Hence, the joint probability of words $\boldsymbol{w}$, topics $\boldsymbol{z}$, timestamps $\boldsymbol{t}$, and authors $\boldsymbol{x}$ is

$$P(\mathbf{x},\mathbf{z},\mathbf{w},\mathbf{t}|\Theta,\Phi,\Psi,\mathbf{a}) = \prod_{d=1}^{D}\prod_{i=1}^{N_d}\frac{1}{A_d}\mathrm{Beta}\left(\psi_{z_{di}}^{t_{di}}\right) \times \prod_{z_j=1}^{T}\prod_{z_k=1}^{T}\prod_{v=1}^{V}\prod_{x=1}^{A}\theta_{z_k|z_jx_k}^{m_{z_k|z_jx}}\phi_{z_kv}^{n_{z_kv}} \tag{8}$$

where z_j is the sampled topic for the previous word $\boldsymbol{w}_j$ and z_k is the sampled topic for the current word $\boldsymbol{w}_k$.

By integrating out the variables Θ and Φ, we have

$$\begin{aligned} P(\mathbf{x},\mathbf{z},\mathbf{w},\mathbf{t}|\alpha,\beta,\Psi,\mathbf{a}) = {} & \prod_{d=1}^{D}\prod_{i=1}^{N_d}\frac{1}{A_d}\mathrm{Beta}\left(\psi_{z_{di}}^{t_{di}}\right) \\ & \times \prod_{x=1}^{A}\prod_{z_j=1}^{T}\frac{\Gamma\left(\sum_{z_k=1}^{T}\alpha_{z_k}\right)}{\prod_{z_k=1}^{T}\Gamma\left(\alpha_{z_k}\right)}\frac{\prod_{z_k=1}^{T}\Gamma\left(m_{z_k|z_jx}+\alpha_{z_k}\right)}{\Gamma\left(\sum_{z_k=1}^{T}\left(m_{z_k|z_jx}+\alpha_{z_k}\right)\right)} \\ & \times \prod_{z=1}^{T}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma(n_{zv}+\beta_v)}{\Gamma\left(\sum_{v=1}^{V}(n_{zv}+\beta_v)\right)}. \end{aligned} \tag{9}$$

For the random variables, we estimate Φ using Eq. (4) and Θ by

$$\theta_{z_k|z_jx} = \frac{m_{z_k|z_jx}+\alpha_{z_k}}{\sum_{z'=1}^{T}\left(m_{z'|z_jx}+\alpha_{z'}\right)}. \tag{10}$$

We also use Eqs. (6–7) to estimate ψ_z.

4. Parameter estimation

Exact inference on LDA-style models is an intractable problem. A variety of algorithms have been used for approximate inference, for example variational EM methods [5], Gibbs sampling [11] [32], and expectation propagation [11] [24]. We chose Gibbs sampling for its ease of implementation.

As for the hyperparameters α and β, previous sampling-based treatments usually take a fixed values (e.g., $\alpha = 50/T$ and $\beta = 0.1$). However, we found that in our setting, the models are sensitive to the hyperparameters. We thus used an expectation–maximization (EM) algorithm to find the optimal values of the hyperparameters by maximizing the Eqs. (3) or (9). This results in a Gibbs EM algorithm [2].

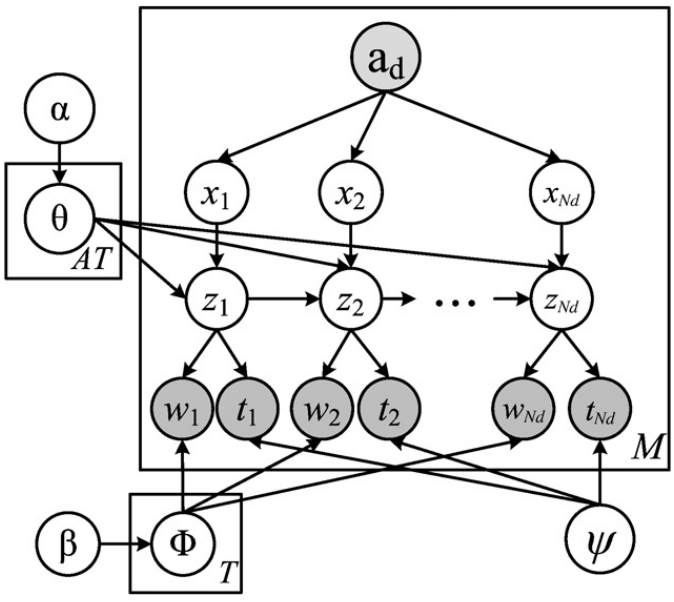


Fig. 2. Graphical model of HMATT.

The algorithm for the parameter estimation in the ATT model can be summarized as follows

1. Initialize $\mathbf{z}$, α, and β.
2. Do until termination
   (a) E-step: for each word token w_{di} in each document d, draw an author x_{di} and a topic z_{di} using a Gibbs sampler.
   (b) M-step: update ψ using Eqs. (6–7) and find the optimal hyperparameters by maximizing

$$argmax_{(\alpha,\beta)} logP(\mathbf{x},\mathbf{z},\mathbf{w},\mathbf{t}|\alpha,\beta,\Psi,\mathbf{a}).$$

Here, the probability in the M-step is defined by Eq. (3). For the HMATT model, we simply replace the E-step by drawing an author–topic pair (x_{di}, z_{di}) using a Gibbs sampler and replace the probability in the M-step by Eq. (9).

In the E-step, for sampling a topic for each word token, the posterior probability for the topic and author is (ATT)

$$P(z_{di},x_{di}|\mathbf{z}_{-di},x_{-di},\mathbf{w},\mathbf{t},\alpha,\beta,\Psi)\propto \mathrm{Beta}\left(\psi_{z_{di}}^{t_{di}}\right) \times \frac{n_{z_{di}w_{di}}^{-di}+\beta_{w_{di}}}{\sum_v\left(n_{z_{di}v}^{-di}+\beta_v\right)}\frac{m_{x_{di}z_{di}}^{-di}+\alpha_{z_{di}}}{\sum_z\left(m_{x_{di}z}^{-di}+\alpha_z\right)} \tag{11}$$

or (HMATT)

$$P(z_{di},x_{di}|\mathbf{z}_{-di},\mathbf{x}_{-di},\mathbf{w},\mathbf{t},\alpha,\beta,\Psi)\propto \mathrm{Beta}\left(\psi_{z_{di}}^{t_{di}}\right) \times \frac{n_{z_{di}w_{di}}^{-di}+\beta_{w_{di}}}{\sum_v\left(n_{z_{di}v}^{-di}+\beta_v\right)}\frac{m_{z_{di}|z_{d(i-1)}x_{di}}^{-di}+\alpha_{z_{di}}}{\sum_z\left(m_{z|z_{d(i-1)}x_{di}}^{-di}+\alpha_z\right)} \tag{12}$$

where the superscript $-t$ denotes a quantity, excluding the current instance (the di-th word token).

In the M-step, given the samples z, the optimal α can be computed using fixed-point iteration

$$\alpha_z^{new}=\frac{\alpha_z\left(\sum_d(\varphi(m_{xz}+\alpha_z)-\varphi(\alpha_z))\right)}{\sum_d(\varphi(m_x+\sum_i\alpha_i)-\varphi(\sum_i\alpha_i))} \tag{13}$$

where $\varphi(x)=\frac{dlog\Gamma(x)}{dx}$ and m_{xz} is the number of times topic z has been associated with author x. Similar fixed-point iterations can be used to estimate β [23]. In both the ATT and HMATT models we assume that there are T hyperparameters for α and V for β. We can also consider other prior settings, for example only one α for all topics z or one α for each topic transition $\theta_{z_k|z_j x}$ (thus $T\times T$ hyperparameters for α).

When applying the learned topic model to new documents, we can perform a few Gibbs sampling iterations to obtain the topic for each word in the new document by (ATT as the example)

$$P(z_{d'i},x_{d'i}|\mathbf{z}_{-d'i},\mathbf{x}_{-d'i},\mathbf{w},\mathbf{t},\alpha,\beta,\Psi)\propto \mathrm{Beta}\left(\psi_{z_{d'i}}^{t_{d'i}}\right) \times \frac{n_{z_{di}w_{di}}+n_{z_{d'i}w_{d'i}}^{-d'i}+\beta_{w_{d'i}}}{\sum_v\left(n_{z_{di}v}+n_{z_{d'i}v}^{-d'i}+\beta_v\right)}\frac{m_{x_{di}z_{di}}^{-di}+\alpha_{z_{di}}}{\sum_z\left(m_{x_{di}z}^{-di}+\alpha_z\right)} \tag{14}$$

where $n_{z_{di}w_{di}}$ is the number learned from the training data and $n_{z_{d'i}w_{d'i}}^{-d'i}$ denotes the number of times of word $w_{d'i}$ occurring in the new document d', excluding the current instance.

4.1. Parallelization

As the Gibbs sampling algorithm for parameter estimation needs to make multiple passes over the entire data set, it often takes multiple days (even weeks) to learn the topic model on a large scale of the scientific literature data, which makes it impractical for many applications. Inspired by the distributed inference for LDA [27], we implement a distributed inference algorithm over multiple processors for the proposed models. We now use the ATT model as example to explain how we perform the parallel training of the topic model. The basic idea is to conduct the inference in a "distribute-and-merge" way. In the distribution step, given P processors, we distribute the document collection D over the P processors, with $D_p=D/P$ documents on each processor. Then we partition the author-specific (author by topic) count matrix to the P processors and duplicate the other (topic by word, topic by time) matrices to each processor. For parameter estimation, we conduct Gibbs sampling on each processor for the

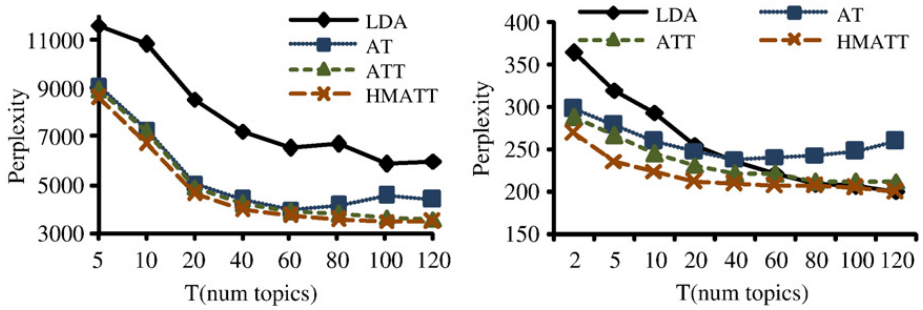


Fig. 3. Average perplexity of five-fold cross-validation obtained by LDA, AT, ATT, and HMATT on the two data sets: Yahoo Newsgroup (left) and NIPS papers (right).
Fig. 4. Evolution of five topics with time on NEWSGROUP (HMATT). For a better view of the trend, we use a square root function for each topic probability here.

distributed documents for a number of internal iterations independently. In the internal iteration, the duplicated matrices will be updated independently. Essentially, for performing a fixed number of sampling iterations, we would like to have a large number of the internal iterations, accordingly can have a small number of distribution-and-merge steps, which will reduce the communication cost of duplicating the matrices. However, since the internal iteration updates the duplicated matrices independently, a large number of internal iterations will result in an incorrect result. Thus, we need a trade-off between the communication cost and the correctness. In our experiments, we empirically increase the number from 1 and test the difference in the perplexity of the parallel ATT model with the single-machine version, and finally select 5 as the number of internal iteration. In the merging step, we combine the count matrices to guarantee the consistency of the count matrices. More specifically, we respectively update each element of two duplicated (topic by word, topic by conference) matrices by

$$n_{zw}^{(\text{new})} = n_{zw}^{(\text{old})} + \sum_{p=1}^{P} \left(n_{zw}^{(p)} - n_{zw}^{(\text{old})}\right) \tag{15}$$

$$t_{z}^{y(\text{new})} = t_{z}^{y(\text{old})} + \sum_{p=1}^{P} \left(t_{z}^{y(p)} - t_{z}^{y(\text{old})}\right) \tag{16}$$

$$s_{z}^{2(\text{new})} = s_{z}^{2(\text{old})} + \sum_{p=1}^{P} \left(s_{z}^{2(p)} - s_{z}^{2(\text{old})}\right) \tag{17}$$

where the number $n^{(\text{old})}$ with the superscript (old) denotes the count before distribution and the number $n^{(\text{new})}$ with the superscript (new) denotes the count after merging. The number $n^{(p)}$ denotes the count obtained after the independent sampling on each processor. The distributed inference algorithm can be considered as an approximation of the single-processor inference algorithm Experimental results in Section 5 show that such approximation can obtain effective results. So far, the distributed training algorithm has been implemented using Hadoop.[1]

4.2. Computational complexity

We analyze the complexity of the proposed topic models. The ATT model has a complexity of $O(MD\overline{N}_dT)$, where M is the number of sampling iterations, and N_d is the average number of word tokens in a paper. The HMATT model has a higher complexity $O(MDA\overline{N}_dT)$. In the parallel ATT model, the time complexity is $O(M((D/P)D\overline{N}_dT) + (M/I_p)(TV))$, where I_p is the number of internal iterations on each processor. $(M/I_p)(TV)$ is the time complexity of duplicating and merging the matrices to/from each processor. Similarly, the complexity of the parallel HMATT is $O(M((D/P)DA\overline{N}_dT) + (M/I_p)(TV))$. We see that with the parallelization over multiple processors (e.g., 100 processor) and with an appropriate number of internal iterations (e.g., 10), we can obtain a significant reduction of the time complexity.

5. Experimental results

5.1. Experiment setting

We conducted experiments on two real-world data sets: NIPS conference papers[2] and NEWSGROUP posts (emails from two Yahoo groups).

The NIPS data set [10] consists of the full text of the 12 years of proceedings, from 1988 to 1999, of the Neural Information Proceedings Systems (NIPS) conferences. The data set contains $D = 1740$ papers with $A = 2442$ authors. For the NEWSGROUP data

[1] http://hadoop.apache.org/core/.
[2] http://www.cs.toronto.edu/ roweis/data.html.

Table 2
Accuracy of time prediction (%).

	LDA	AT	ATT	HMATT
NIPS	40.75	43.81	84.37	86.46
NEWSGROUP	51.57	55.90	88.01	89.24

set, we randomly chose $D = 1218$ posts from the NBA2DAY and JENA groups of Yahoo from 2004 to 2005. These posts were authored by $A = 216$ unique authors (each post has one author). The author information of NIPS is extracted from the paper's metadata information and the authors in NEWSGROUP are the owners of the posts. We do not apply advanced NLP techniques (such as named entity recognition) to the data set, and simply preprocess each data set by (a) removing stopwords and numbers; (b) removing words that appear less than three times in the corpus; and (c) downcasing the obtained words. Finally, we obtained 28,928 unique words and a total of 2,985,728 word tokens in the NIPS data set and 8223 unique words and a total of 180,500 word tokens in the NEWSGROUP data set. In NIPS, each document's timestamp is determined by the year of the proceedings and in NEWSGROUP; each post's timestamp is determined by the posted time (with format "YYYY-MM-DD Hour:Min:Sec").

In the rest of the section, we first compare the proposed models with Latent Dirichlet Allocation (LDA) and the Author–Topic (AT) model. Then we analyze the results obtained by our models on the two data sets.

5.2. Perplexity and time prediction

We evaluated the performances of the proposed models and the LDA and AT models in terms of *Perplexity*, a standard measure for estimating the performance of a probabilistic model. The perplexity of an unseen test document $(\boldsymbol{w}_d, \boldsymbol{a}_d) \in D^{\text{test}}$ is defined as

$$Perplexity(\mathbf{w}_d \mid \mathbf{a}_d) = exp\left(-\frac{P(\mathbf{w}_d \mid \mathbf{a}_d)}{N_d}\right). \tag{18}$$

Table 3
Three topics discovered by ATT (above) and HMATT (bottom) for the NEWSGROUP data set. Each topic is shown with the top 10 words and their corresponding conditional probabilities. Below are top 6 authors associated with each topic. The titles are our interpretation of the topics.

Topic #4 (ATT) "Yao at preseason"		Topic #8 (ATT) "RDQL querying"		Topic #9 (ATT) "Ontology reasoning"	
Camp	0.017670	Query	0.050047	Reason	0.039988
Yao	0.013837	RDQL	0.038735	File	0.024535
Preseason	0.010689	Type	0.026921	ABox	0.022274
Pippen	0.009594	OWL	0.018123	Data	0.018505
Million	0.008910	Result	0.017118	TBox	0.018128
China	0.007678	id	0.013850	Question	0.016998
Basketball	0.007268	RDF	0.012593	Model	0.016244
NET	0.006583	Namespace	0.011337	Owl:class	0.013606
Contract	0.006446	kn:id	0.010080	Ontmodelspec	0.012852
Training	0.005899	Jena	0.009074	Transition	0.011721
GarysLists	0.842239	Chris Dollin	0.465006	Andrew_crapo	0.667318
Alper Kuray	0.004627	Seaborne	0.104389	Ignazio Palmisano	0.040059
GaryBayside	0.003134	Sven Abels	0.055160	Ellis R Watkins	0.028334
Scott Davis	0.002090	Abelssoft	0.050119	Howard Goldberg	0.027846
Gary Gentile	0.000781	David Vallet	0.049822	Dave Reynolds	0.023449
Yorin15	0.001642	Andy	0.047450	Alessandro Di Bella	0.005374
Topic #44 (HMATT)		Topic #11 (HMATT)		Topic #1 (HMATT)	
Camp	0.018086	RDQL	0.032128	Reason	0.025874
Preseason	0.012784	Query	0.014283	ABox	0.015941
Yao	0.011998	Property	0.012290	TBox	0.014995
Train	0.008856	Subject	0.009959	File	0.011841
China	0.007678	Listsubject	0.009298	Model	0.011683
Exhibit	0.007482	Jena	0.006981	Statement	0.011210
Shanghai	0.007285	Type	0.006651	Problem	0.010580
Pippen	0.005518	RDF	0.006320	Query	0.010264
NBA	0.005322	Creator	0.006320	Classpath	0.010107
Scout	0.005125	Null	0.005989	Spec	0.009318
GarysLists	0.875649	Soledad Calo	0.295944	Alessandro Di Bella	0.239888
Alper Kuray	0.001238	Chris Dollin	0.204967	Andrew_crapo	0.237099
Bounce mybuns12	0.001009	Seaborne	0.103311	Chris Dollin	0.219491
Gary Gentile	0.000781	Abelssoft	0.072490	Dave Reynolds	0.020921
Yorin15	0.000552	Ian Dickinson	0.022483	Elias Torres	0.020223
Robert Littal	0.000552	Ignazio Palmisano	0.0067015	Ignazio Palmisano	0.019351

We compute the perplexity of a held-out test set to evaluate each topic model. All the topic models were trained using the same data and with the same Gibbs EM algorithm. Better generalization performance is indicated by a lower perplexity.

We present the five-fold cross-validation results. All topics were extracted at the 2000th iteration of the Gibbs sampler. All experiments were carried out on a Server running Windows 2003 with two Dual-Core Intel Xeon processors (3.0 GHz) and 8 GB memory. It takes 20 min and 50 min, respectively, to estimate the LDA model and the AT model, 60 min for the ATT model and 1.35 h for the HMATT model. When the training is distributed over 8 processors, the speedup is 3.7. That is, for training the ATT and the HMATT models, it takes only 20 min (ATT) and 30 min (HMATT).

Fig. 3 plots the average perplexity of the four models with different numbers of topics. First we find that ATT and HMATT have similar patterns to LDA: the perplexity score first decreases with increasing numbers of topics, and then increases when the number becoming large. We see that on the two data sets, both the ATT and HMATT models clearly outperform LDA and are slightly better than the AT model in terms of perplexity. The AT model outperforms LDA on NEWSGROUP and is better than LDA on NIPS when the number of topic is small; however it underperforms LDA as number increases. We conduct a statistical test. Specifically, we train the model on the training data set (five-fold cross-validation) and calculate the perplexity of the test data set according to the trained model. Taking the number of topics as $T=50$, the p-values of the sign test are 0.00089 and 0.0062 by LDA and AT on Yahoo and 0.0089 and 0.0074 on NIPS, which indicates that at a significance level of 0.01, the improvements of our models over the baseline models are statistically significant.

We also performed an experiment of time prediction on the two data sets, that is, predicting the timestamp for the given document. We used accuracy as the evaluation measure, and compared our models with different topic models. We want to show whether a model that simultaneously models the evolution of document contents and author interests can improve the accuracy of prediction. Specifically, on NIPS, given a paper with authors, we predict its publication year, and on NEWSGROUP, given a post with an author, we predict its posting year and month. We predict the timestamp for each document by choosing the discretized timestamp that maximizes the posterior, which is calculated by multiplying the timestamp probability of all word tokens, e.g., in ATT and HMATT, we have $argmax_t \Pi_{i=1}^{N_d} p(t|\psi_{z_i})$. In the LDA and AT models, we use the estimated model and the associated timestamp to each document to obtain the timestamp by maximizing $argmax_t \Pi_{i=1}^{N_d} p(t|z_i)$.

Table 4
Three topics discovered by ATT (above) and HMATT (bottom) for the NIPS data set.

Topic #9 (ATT) "Support vector machine"		Topic #22 (ATT) "Speech recognition"		Topic #37 (ATT) "Neural network"	
Distance	0.026854	Speech	0.051015	Network	0.067485
Kernel	0.027055	Word	0.040950	Dynamic	0.042527
Vector	0.027714	Phoneme	0.030456	Neural	0.031825
Set	0.017547	Speaker	0.020019	System	0.025906
Method	0.013402	Recognition	0.018532	State	0.019793
Pattern	0.012259	Perform	0.013728	Neuron	0.019583
Support	0.012031	Acoustic	0.009725	Equation	0.019373
Machine	0.010090	Test	0.009668	Point	0.016115
Train	0.009291	Train	0.009611	Connect	0.012770
Transform	0.009090	Phonetic	0.009210	Function	0.010966
Patrice Simard	0.068918	Alex Waibel	0.114568	O.J. Pineda	0.019984
Alex Smola	0.059160	Ajay Jain	0.028099	R.M. Westervelt	0.019966
Vladimir Vapnik	0.035141	Ronald Cole	0.024188	Todd K. Leen	0.019349
Holger Schwenk	0.033370	Candace Kamm	0.021002	Andre Longtin	0.014741
Trevor Hastie	0.027068	Victor Zue	0.018742	Frank Eeckman	0.014088
Jonathan Baxter	0.025006	Regis Cardin	0.018395	H. Sebastian Seung	0.013099
Topic #14 (HMATT)		Topic #36 (HMATT)		Topic #2 (HMATT)	
Vector	0.034456	Speech	0.072350	Network	0.120764
Support	0.028355	Word	0.034516	Neural	0.090750
Distance	0.022793	Recognition	0.0290371	Unit	0.060281
Kernel	0.020901	Speaker	0.022783	Input	0.050072
Tangent	0.015369	HMM	0.019625	Train	0.040774
Pattern	0.014881	Perform	0.011329	Output	0.036331
Machine	0.010782	Acoustic	0.010623	Learn	0.029981
Error	0.010753	Phoneme	0.009792	Weight	0.028338
Margin	0.009864	System	0.009523	Hidden	0.027907
Vector	0.009291	Phonetic	0.008990	Layer	0.025532
Alex Smola	0.059160	Alex Waibel	0.123784	Michael Mozer	0.017901
Vladimir Vapnik	0.045141	Renato De Mori	0.067076	Geoffrey Hinton	0.013248
Chris Burges	0.034678	Ronald Cole	0.042206	Dean Pomerleau	0.011453
Nello Cristianini	0.022169	Khalid Choukri	0.027352	Terrence Sejnowski	0.011175
Trevor Hastie	0.017031	Candace Kamm	0.023007	Alex Waibel	0.010990
John Denker	0.013989	Regis Cardin	0.010139	Paul Munro	0.009571

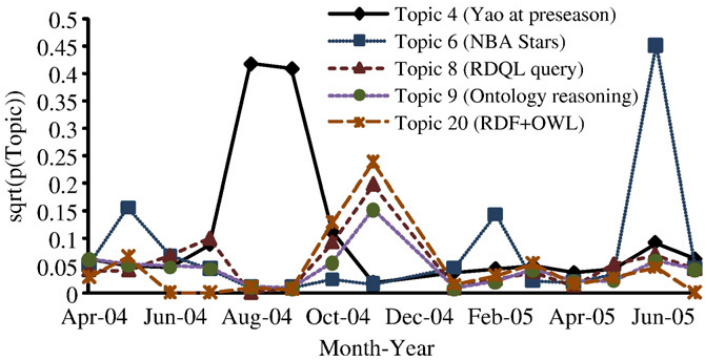


Fig. 4. Evolution of five topics with time on NEWSGROUP (HMATT). For a better view of the trend, we use a square root function for each topic probability here.

Table 2 shows the five-fold cross-validation results of the time prediction for the two data sets. We see that ATT and HMATT clearly outperform the LDA and AT models. Again, HMATT outperforms ATT by +2.09% and +1.23% in terms of accuracy.

5.3. Analysis

We analyzed the results of ATT and HMATT on the two data sets. For simplicity, for all experiments in this section we fix the number of topics ($T=50$).

Tables 3 and 4, respectively, show three topics on the two data sets obtained by the ATT and HMATT models, with aligning by the KL divergence of the topics. Figs. 4 and 5 plot the occurrence probability of five topics, respectively, from NEWSGROUP and NIPS at different times, using HMATT.

By comparing the two figures, we can see several interesting results. The NEWSGROUP data shows stronger temporal patterns. Many of the topics found have sharply shaped trends. Some topics (e.g., Topic #4 and Topic #6) from NEWSGROUP are very sensitive to the time. These results have a clear explanation. Topic #4 talks about "Yao at preseason". As the preseason of the NBA is usually from Sep. to Oct., the topic comes into prominence sharply at that time and becomes silent very quickly when the

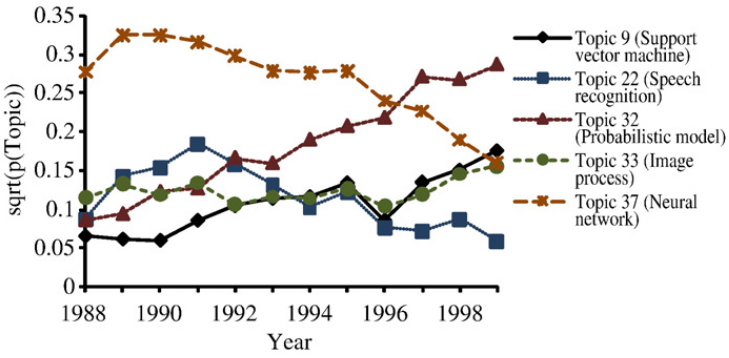


Fig. 5. Evolution of five topics on NIPS (HMATT).

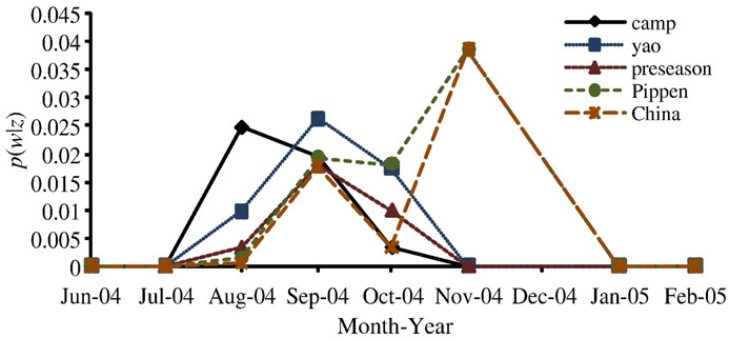


Fig. 6. Evolution of representative words in the topic "Yao at preseason" (HMATT).

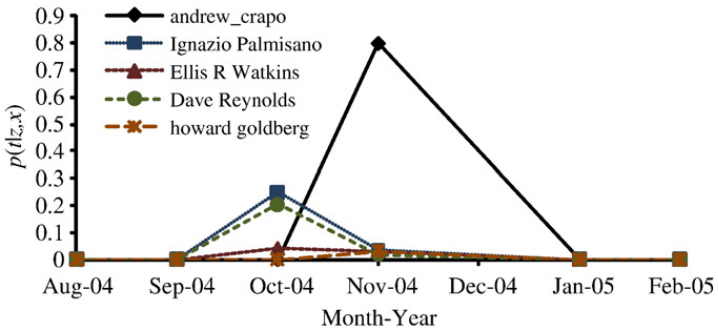


Fig. 7. Evolution of five representative authors in the topic "Ontology reasoning" (HMATT).

preseason is over. The reason that the word "Yao" is hot in the topic is that Chinese player Ming Yao became a highlight in the NBA in 2004 and obtained great achievements in that season.

Topics discovered from the NIPS data are relatively more stable, as research work often spans a longer period from its rise to fall in prominence. We can still see some interesting trends: neural networks become popular at the beginning of 90s and taper off slowly after that; while probabilistic models and support vector machines have become more popular in recent years. In the following analysis, we focus on the analysis of the NEWSGROUP data set due to its stronger temporal patterns and also because there is already much research analyzing the NIPS data, e.g., [31] and [40].

Fig. 6 plots the trends of changes in a few representative terms in the topic "Yao at preseason" by HMATT. It is very interesting that the words "camp", "Yao" exhibit higher emission rates in the period of the NBA preseason (from Aug. to Oct.) and taper off quickly with the end of the preseason (Nov.). The words "Pippen" and "China" quickly come into prominence in Oct. 2004 because there were many discussions about Pippen's retirement after the preseason and there was a match played in China.

Fig. 7 plots the changes in interests of five representative authors on the topic "Ontology reasoning", by HMATT. We can again see very sharply shaped changes. It seems that people usually like to post something for intensive discussion on the newsgroup and quickly become silent or switch to another topic of discussion.

In addition, we plot changes in the topic interest of the author "GarysLists" in five different topics by using HMATT, as shown in Fig. 8. We see on NEWSGROUP, the author's interest may change quickly on some topic (e.g., Topic #40). There are also some authors who seem to have stable interests, for example, from Apr. 2004 to Jan. 2005, the author "milicic9" posted different kinds of messages on Topic #10 and seldom posted messages on other topics.

6. Conclusion

In this paper, we have investigated the problem of how to model trends of changes in the associated data simultaneously. We have proposed two generative models, i.e., Author–Time–Topic (ATT) model and Hidden Markov Author–Time–Topic (HMATT) model, to perform the task. We have used Gibbs EM for approximate inferences. Experiments show that the proposed models clearly outperform the state-of-the-art topic models, the LDA and the Author–Topic models in terms of perplexity and time prediction. Analysis on the topic finding results also unveils some interesting patterns in the topics.

There are many potential future directions for this work. It would be interesting to investigate the regularization method (e.g., [22]) for modeling the evolution of the associated data more smoothly. It would also be interesting to model the citation information in the evolution model (e.g., [8]). In addition, it would also be interesting to apply the method to other domains, for example modeling the evolution of image and associated tags.

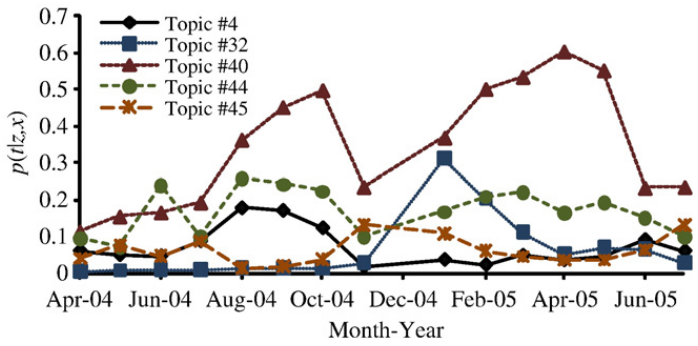


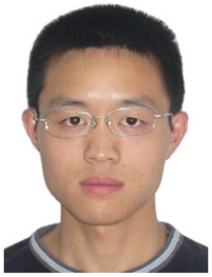
Fig. 8. Interest change of the author "GarysLists" in five different topics (HMATT).

Acknowledgements

The work is supported by the Natural Science Foundation of China (No. 60703059, No. 60973102), the Chinese National Key Foundation Research (No. 60933013), the National High-Tech R&D Program (No. 2009AA01Z138), and the Chinese Young Faculty Research Fund (No. 20070003093).

References

[1] A. Ahmed, E. P. Xing, W. W. Cohen, and R. F. Murphy. Structured correspondence topic models for mining captioned figures in biological literature. In Proceeding of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'09).
[2] C. Andrieu, N. de Freitas, A. Doucet, M.I. Jordan, An introduction to MCMC for machine learning, Mach. Learn. 50 (2003) 5–43.
[3] D.M. Blei, J.D. Lafferty, Correlated topic models, Proceedings of the 18th Neural Information Processing Systems (NIPS'06), 2006.
[4] D.M. Blei, J.D. Lafferty, Dynamic topic models, Proceedings of the 23th International Conference on Machine Learning (ICML'06), 2006, pp. 113–120.
[5] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[6] J. Boyd-Graber, D.M. Blei, Syntactic topic models, Proceedings of the 20th Neural Information Processing Systems (NIPS'08), 2008.
[7] K. Deschacht, M.F. Moens, Semi-supervised semantic role labeling using the Latent Words Language Model, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09), 2009, pp. 21–29.
[8] L. Dietz, S. Bickel, T. Scheffer, Unsupervised prediction of citation influences, Proceedings of the 24th International Conference on Machine Learning (ICML'07), 2007, pp. 233–240.
[9] G. Doyle, C. Elkan, Accounting for burstiness in topic models, Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09), 2009, pp. 281–288.
[10] A. Globerson, G. Chechik, F. Pereira, N. Tishby, Euclidean embedding of co-occurrence data, J. Mach. Learn. Res. 8 (2007) 2265–2295.
[11] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of PNAS'2004, 2004, pp. 5228–5235.
[12] T.L. Griffiths, M. Steyvers, D.M. Blei, B. J., Integrating topics and syntax, Proceedings of the 16th Neural Information Processing Systems (NIPS'04), 2004.
[13] A. Gruber, M. Rosen-Zvi, Y. Weiss, A. Gruber, M. Rosen-Zvi, Y. Weiss, Hidden topic Markov models, Proceedings of Artificial Intelligence and Statistics (AISTATS'07), 2007.
[14] A. Gruber, M.R. Zvi, Y. Weiss, Latent topic models for hypertext, Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI'2008), 2008, pp. 352–359.
[15] T. Hofmann, Probabilistic latent semantic indexing, Proceedings of the 22th ACM SIGIR International Conference on Information Retrieval (SIGIR'99), 1999, pp. 50–57.
[16] T. Iwata, T. Yamada, N. Ueda, Modeling social annotation data with content relevance using a topic model, Proceedings of the 21st Neural Information Processing Systems (NIPS'09), 2009.
[17] J. Leskovec, L. Backstrom, J. Kleinberg, Meme-tracking and the dynamics of the news cycle, Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09), 2009, pp. 497–506.
[18] Y. Liu, A. Niculescu-Mizil, W. Gryc, Topic-link LDA: joint models of topic and author community, Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09), 2009, pp. 665–672.
[19] A. McCallum, Multi-label text classification with a mixture model trained by EM, Proc. of AAAI'99 Workshop on Text Learning, 1999.
[20] A. McCallum, X. Wang, A. Corrada-Emmanuel, Topic and role discovery in social networks with experiments on Enron and academic email, J. Artif. Intell. Res. (JAIR) 30 (2007) 249–272.
[21] S. McClean, B. Scotney, P. Morrow, K. Greer, Knowledge discovery by probabilistic clustering of distributed databases, Data Knowl. Eng. 54 (2) (2005) 189–210.
[22] Q. Mei, D. Cai, D. Zhang, C. Zhai, Topic modeling with network regularization, Proceedings of the 17th International Conference on World Wide Web (WWW'08), 2008, pp. 101–110.
[23] T. Minka. Estimating a dirichlet distribution. In Technique Report, http://research.microsoft.com/ minka/papers/dirichlet/, 2003.
[24] T. Minka, J. Lafferty, Expectation–propagation for the generative aspect model, Proceedings of the 18st Conference on Uncertainty in Artificial Intelligence (UAI'02), 2002, pp. 352–359.
[25] R.M. Nallapati, A. Ahmed, E. Xing, W.W. Cohen, Joint latent topic models for text and citations, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2008), 2008, pp. 542–550.
[26] R.M. Nallapati, S. Ditmore, J.D. Lafferty, K. Ung, Multiscale topic tomography, Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07), 2007, pp. 520–529.
[27] D. Newman, A. Asuncion, P. Smyth, M. Welling, Distributed inference for latent Dirichlet allocation, Proceedings of the 19th Neural Information Processing Systems (NIPS'07), 2007.
[28] U. Nodelman, C.R. Shelton, D. Koller, Continuous time Bayesian networks, Proceedings of the 18st Conference on Uncertainty in Artificial Intelligence (UAI'02), 2002, pp. 378–387.
[29] D.P. Putthividhya, H.T. Attias, S. Nagarajan, Independent factor topic models, Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09), 2009, pp. 833–840.
[30] G. Quon, Y.W. Teh, E. Chan, T.R. Hughes, M. Brudno, Q. Morris, A mixture model for the evolution of gene expression in non-homogeneous datasets, Proceedings of the 20th Neural Information Processing Systems (NIPS'08), 2008, pp. 1297–1304.
[31] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author–topic model for authors and documents, Proceedings of the 20st Conference on Uncertainty in Artificial Intelligence (UAI'04), 2004.
[32] M. Steyvers, P. Smyth, T. Griffiths, Probabilistic author–topic models for information discovery, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), 2004.
[33] J. Tang, H.f. Leung, Q. Luo, D. Chen, J. Gong, Towards ontology learning from folksonomies, Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09), 2009, pp. 2089–2094.
[34] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: extraction and mining of academic social networks, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'08), 2008, pp. 990–998.
[35] J. Tang, J. Zhang, J.X. Yu, Z. Yang, K. Cai, R. Ma, L. Zhang, Z. Su, Topic distributions over links on web, Proceedings of 2009 IEEE International Conference on Data Mining (ICDM'09), 2009.
[36] J. Van Hulse, T. Khoshgoftaar, Knowledge discovery from imbalanced and noisy data, Data Knowl. Eng. 68 (12) (2009) 1513–1542.
[37] E. Štrumbelj, I. Kononenko, M. Robnik Šikonja, Explaining instance classifications with interactions of subsets of feature values, Data Knowl. Eng. 68 (10) (2009) 886–904.
[38] H.M. Wallach, Topic modeling: beyond bag-of-words, Proceedings of the 23rd International Conference on Machine Learning (ICML'06), 2006, pp. 977–984.
[39] W. Wang, B. Payam, B. Andrzej, Probabilistic topic models for learning terminological ontologies, IEEE Trans. Knowl. Data Eng. (TKDE) (2009).
[40] X. Wang, A. McCallum, Topics over time: a non-Markov continuous-time model of topical trends, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge discovery and Data Mining (KDD'06), 2006, pp. 424–433.
[41] Q. Zhao, T.-Y. Liu, S.S. Bhowmick, W.-Y. Ma, Event detection from evolution of click-through data, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06), 2006, pp. 484–493.



Jie Tang received the PhD degree from Tsinghua University. He is an associate professor in Department of Computer Science and Technology of Tsinghua University. His research interests include machine learning, text mining, social network analysis, and semantic web.

Jing Zhang received the MS degree in computer science from Tsinghua University in 2008. Her research interests include information retrieval and text mining.