International Journal of Database Theory and Application
Vol.8, No.1 (2015), pp.213-226
http://dx.doi.org/10.14257/ijdta.2015.8.1.22

Time Label Topic Model

YongHeng Chen[1], Wanli Zuo[2], kerui Chen[3] and Yaojin lin[4]

[1,4]College of Computer Science, Minnan Normal University, zhangzhou 363000, China
[2]College of Computer Science and Technology, Jilin University, Changchun, China
[3]College of Computer and Information Engineering, Hennan University of Economics and Law, Zhengzhou
Response author:Yongheng Chen email: cyh77!@163.com

Abstract

Most of the models not aware of these dependencies on document time stamps. Not modeling time can confound co-occurrence patters and results in exchangeability of topic problem, which is important factor to deal with when finding dynamic topic discovery. This limitation has thus motivated work on developing a generalized framework for incorporating time information into topic models. Consequently, a topic model named Topics over Time (TOT) is proposed, which introduces a time node in topic model to handle the exchangeability of topics problem. However it lacks the capability to accommodate data type of side information. In this paper, we present a generative time LDA-style topic model with a variety of side information named Time Label Topic(TLT), which can find not only how the latent low-dimensional structure of document-response pairs changes over time, but also overcome the exchangeability of topics problem. Empirical results demonstrate significant improvements accuracy of time stamp and response variable prediction, and lower perplexity of our proposed model and dominance over other models.

Keywords: *Bayesian Models, Gibbs Sampling, Variational Expectation-Maximization, Topic Modeling*

1. Introduction

As peoples began memorizing their documents in digital, Information Retrieval(IR) as a domain rose in computer science. Information Retrieval increased with the wide extend application of web technology. Peoples need find relevant web pages that are satisfying their information need from millions of web pages on the internet through a convenient and efficient way. Describing the feathers of documents' content is a typical problem emphasized in Information Retrieval. Researchers usually employ the characteristic of the content of document to search, form, or classify the corpus.

Recently, generative models for corpus have been used to detect topic-based content presentations, each documents are modeled as a mixture of probabilistic topics. A statistical generative model, such as the Probabilistic Latent Semantic Indexing (PLSI), is proposed by Hofmann, which is one of topic model and employs topics represented by latent variables to connect documents and words [1]. A document is regarded as a mixture of topics. The content, the words in a document, can been produced presented the small set of topics (or latent variables). Reversing this process, *i.e.*, matching the generative model to the words in training set, equivalent to deducing the latent variables and, therefore, Inferring the potential topics' distributions. Blei *et al.*, proposed Latent Dirichlet Allocation (LDA), which develops the

ISSN: 2005-4270 IJDTA

generative model to accomplish the ability of concluding generalizing the topic distributions so that research can also employ LDA model to create unseen document [2]. The achieving success of Latent Dirichlet Allocation in the research field far exceeds the domain of Information Retrieval. There has been a wide range of applications employed LDA model in relevant area, for instance, multimedia classification and data mining.

Although Latent Dirichlet Allocation model has sufficiently capacity to acquire the topics distribution for a document, as an unsupervised model, this model cannot provide evidently approach of incorporating a supervised label into model's learning analysis. In order to incorporate supervised label, some alterations of Latent Dirichlet Allocation are been put forward in existing literature. The supervised Latent Dirichlet Allocation (sLDA) model emphasizes the prediction issue through deducing the most predictive potential topics of labeled document [6]. The Dirichlet-multinomial regression (DMR) topic model is put forward by Mimno *et al.*, which includes a log-linear prior on the document-topic distributions, where the prior is a function of the observed document features [7]. The essential difference between Dirichlet-multinomial regression and supervised Latent Dirichlet Allocation is that, while supervised Latent Dirichlet Allocation model regard observed characteristic as generated variables, Dirichlet-multinomial regression treats the observed characteristic as a set of conditioned variables.

However none of the above-mentioned topic models take into account the dependencies on document time stamps. Applying model individually for each documents separated by time stamp during acquiring the topics of document will give rise to exchangeability of topics problem in general, which implies that a topic model in disparate operates will not have alike topics and the sequence of topics will not be the same as well. Then a topic modeling method, Topics over Time (TOT), was proposed by wang *et al.*, which gains the evolution of document's topics by incorporating a time node into LDA model to resolve the exchangeability of topics problem, whereas this model leave the label of document out of consideration [5].

In this paper, Time and Label Topic (TLT), a topic modeling approach, is presented, which considers the response variable associated with document, jointly with continuous time stamps. TLT model classifies discovered topics by the response variable, predicts the response variables and time stamps for unseen documents, and tracks dynamic evolution of the discovered topics over time.

To my best of our knowledge, we are the first to deal with the dynamic thematic pattern' analysis with respect to topics by dependently modeling all time slices simultaneously based on response variables attached with documents.

The outline of this paper is organized as follows. In the next Section, our Time Label Topic model is introduced. In Section 3, Gibbs sampling and Variation Expectation-Maximization are utilized for parameter estimation. In Section 4, the method of classifying the discovered topics and the definition of the attention degree of topic are given. We present the experiments we carry out on the basis of data based on Xinhua News corpus and the results we obtained in Section 5. Our final conclusions and suggestions for future work are discussed in Section 6.

2. Time Label Topic

2.1 Modeling Documents with Topics

Before presenting the Time Label Topic (TLT) model, let us review the basic Latent Dirichlet Allocation model. A glossary of notations used in the paper is summarized in Table 1, and the graphical model representations of our TLT model is shown in Figure 1.

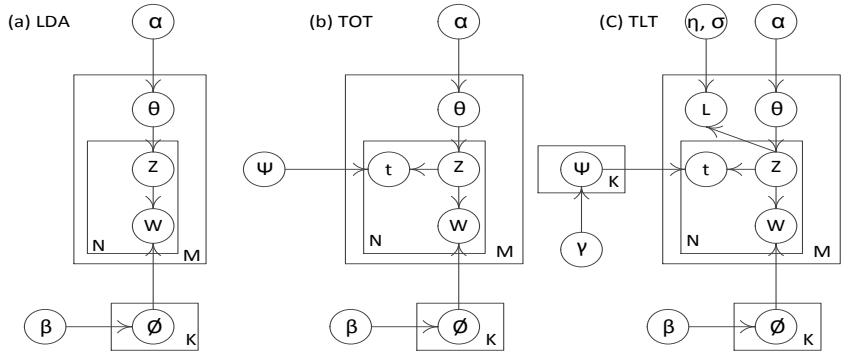

Figure 1. TLT is Shown with Two Inspiration Models: (a) LDA Model (b) TOT Model (c) TLT Topic Model

Table 1. Notation used in the Paper

SYMBOL	DESCRIPTION
K	number of topics
M	Total number of documents
V	Total number of unique words
N_d	number of word tokens in document d
α_d	K-vector of dirichlet priors for ducument d
β_k	V-vector of dirichlet priors for topic k
γ_k	symmetric dirichlet priors for Ψ_k
θ_d	the multinomial distribution of topics specific to the document d
$\emptyset_k$	the multinomial distribution of words specific to the topic k
Ψ_k	the multinomial distribution of words specific to the topic k with a time stamp
σ η	The response parameters σ and η
L_d	The response of document d
Z_{di}	the topic associated with the ith token in the document d
W_{di}	the ith token in document d
t_{di}	the timestamp associated with the ith token in the document d

Latent Dirichlet Allocation model, Figure 1 (a), is a hierarchical Bayesian network that product a document making use of a mixture of topics. The parameters θ and Ø, the topic and document distributions are conditionally separation because of the words that can be observed. Moreover, the direct connection between documents and words is interrupted. Rather, by introducing additional potential variable z, which represents the responsibility of an especial topic in employing that word in the document, *i.e.*, the topic(s) that the document is centralized on, the connection is managed. The generative process of the topic model specifies a probabilistic sampling procedure that describes how words in documents can be generated based on the hidden topics. The document and topic distributions adopt respectively α and β as the Dirichlet priors. To deal with unseen documents the generative model of LDA is generalized. For LDA model the words in a document is exchangeable, this

is the same for the documents in a corpus. The generative process of the topic model specifies a probabilistic sampling procedure that describes how words in documents can be generated based on the hidden topics [9].

LDA produced documents through selecting a distribution on topics θ from a Dirichlet distribution. P (z) is determined by θ for words in document. The words in the document are then generated by picking a topic j from this distribution and then picking a word from that topic according to P(w|z=k), which is determined by a fixed Øk [3]. The estimation problem becomes one of maximizing P (w|Ø, α)=∫P(w|θ, Ø)P(θ|a)dθ,where P(θ) is a Dirichlet(a) distribution. θ and Ø are usually estimated by using sophisticated approximations, either Gibbs sampling or Variational Expectation-Maximization [10].

Whereas, Latent Dirichlet Allocation model neglected the natural term changed by time discretization, which is important to acquire the trend of topics [11]. Then Topics over Time model was presented. Word co-occurrences and temporal information affect the detection of topic in TOT model. Rather than modeling a sequence of state shifts with a Markov assumption on the dynamics, TOT models (normalized) absolute time stamp values [8].

Considering a document that is not labeled with time stamp, TOT can predict absolute time values for it by study long-range dependencies in time. And TOT model also predict topic distributions based on a time stamp. Markov model's risk of inappropriately dividing a topic in two when there is brief gap in its appearance can be prevented [5]. The topic mixture multinomial distribution θd is sampled from Dirichlet prior parameter α for each document in TOT model. And then the potential topic z is selected and a word w with time stamp t is produced from topics-specific multinomial distributions Øk and beta distribution Ψk , respectively, over words and time stamp of a document for that topic (see Figure 1(b)) [12].

2.2. Time Label Topic Model

While TOT is expressional plenty to reveal model's topics of documents related with a sequential distribution over time stamps, and represents great for predicting time stamp and evolution, it is not adequate for modeling data set paired with a response, or labeled, since, as an unsupervised model, it provides unclear mean of combining response variables into model's training procedure. This simulated us to extend TOT model to combine and handle differently response types of data set. Through introducing generalize linear models, preferred Method of sLDA model, to a changed TOT model, we propose TLT model (see Figure 1(c)), which can find how the latent low-dimensional structure of document-response pairs and analyze its changing over time.

In our approach, we view a document associated with only one response variable and time stamp as a mixture of topics, and each word in this document having a copied response label and time stamp during generative process on training set. However, after smoothing the model from training data set, if actually run as a generative model, the words within the same document would been produced variously response variables and time stamps. Because the topic is the distribution of words, the response variable of topic can be calculated by the response variable of words, and furthermore a document without response variable can be labeled by its topics' response variable. A time stamp can be predicted given the words within the document.

In TLT model, a document d is related with a multinomial distribution θd over topics and each topic is related with a multinomial distribution Øz over words and multinomial distribution Ψz with a time stamps for each word of a document for that topic. L is response variable of document. σand η is the response parameters. θd, Øz and Ψz have a symmetric Dirichlet prior with hyper parameters α, β and ٢, respectively.

Symbolically, a data set of M documents can be indicated as: M = {(w_1,l_1,t_1), (w_2,l_1,t_2), . . . , (w_d,l_1,t_d)}, where wd is word vector chosen from a vocabulary of size V, l_d is response value of document d, t_d is the time stamp of document d. So, the generating probability of the word w with time stamp t for a document d with response variable L is given as:

$$P(Z,w,L,t|\ \alpha,\ \beta,\ \sigma,\eta,\text{٢})=P(w,Z|\alpha,\ \beta)P(L|Z,\sigma,\eta)P(t|Z,\text{٢})$$

Under the TLT model, each document and response arises from the following generative process:

For each topic k ∈ {1, . . . ,K}:
Generate $Øk=(Øk,1,...,Øk,v)^T$ ~Dir(.|β)
Generate $Ψk=(Ψk,1,...,Ψk,T)^T$~Dir(.|٢)
For each document d:
Generate $θd=(θd,1,...,θd,v)^T$ ~Dir(.|α)
For each word wdi
Generate zdi ~Mult(.|θd)
Generate wdi ~Mult(.|Øzdi)
Generate twdi ~Mult(.|Ψzdi)
Generate response variable $L|z,\sigma^2,\eta \sim N(\eta^T,\sigma^2)$.

T is the number of time slice. Notice the response comes from a normal linear model. The covariates in this model are the (unobserved) empirical frequencies of the topics in the document [6]. The regression coefficients on those frequencies constitute η. Note that a linear model usually includes an intercept term, which amounts to adding a covariate that always equals one [6]. Here, such a term is redundant, because the components of $\bar{z}$ always sum to one [6].

3. Approximate Variational Inference

We have represented the intention behind TLT model and clarified its conceptual benefits over other models. In this section, shown a data set of documents M= {(w_1, l_1, t1), (w_2, l_1, t_2). . . (w_d, l_1, t_d)}, we shift our attention to process for inference and parameter estimation under TLT model, *i.e.*, find parameters α, β and ٢.

Gibbs sampling is a Markov chain Monte Carlo(MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariateprobability distribution (*i.e.*, from the joint probability distribution of two or more random variables), when direct sampling is difficult. Complexity reduction to be done through the Gibbs sampling algorithm let us transform parameter calculation question into a not complicated counting and sampling course. Variational expectation maximization procedure is extensively applied in parameter estimation problems leading to log-likelihood maximization

In this paper, we combine Gibbs sampling and Variational Expectation-Maximization to perform approximate inference. Based on Variational Expectation-Maximization algorithm we adopt Gibbs sampling to realize variational E-step. Through this process, we will obtain parameters θd, the multinomial distribution of topics specific to the document d, Øz, the multinomial distribution of words specific to topic z, and Ψz, the multinomial distribution of time specific to topic z. The M-step will determine the parameters σand η normal distribution by Maximum likelihood method.

E-step:

In this step, parameters σ, η considered as constant are fixed. We wish to find the special distribution of parameters θ, Ø and Ψ that maximize (marginal) likelihood of p ($z,w,l|\alpha,\beta,\text{٢},\sigma^2,\eta$). This Likelihood probability can been decomposed as following:

$$p(z,w,l|\alpha,\beta,\text{٢},\sigma^2,\eta)$$
$$=p(w|z,\beta)p(z|\alpha)p(l|z,\sigma^2,\eta)\int p(t,\Psi|z,\text{٢})d\Psi$$
$$=\int p(w,\text{Ø}|z,\beta)d\text{Ø}\int p(z,\theta|\alpha)d\theta\int p(t,\Psi|z,\text{٢})d\Psi\ p(l|z,\sigma^2,\eta)$$

The probability $p(l|z,\sigma^2,\eta)$ is normal distribution, reflects the interdependency with topics on response variable. The quantity of it can be computed as following:

$$p(l\mid z,\eta,\sigma^2)$$
$$=\prod_{m=1}^{M}\frac{1}{\sqrt{2\pi\sigma^2\ \pi}}\exp\left(-\frac{(l_m-\eta^T\bar{z}_m)^2}{2\sigma^2}\right) \qquad (1)$$

$$\int p(w,\phi\mid z,\beta)d\phi$$
$$=\int p(w\mid z,\beta,\phi)\,p(\phi\mid\beta)d\phi$$
$$=\int\prod_{k=1}^{K}(\prod_{v=1|k,v}^{V}\phi^{n_{k,v}})\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}(\prod_{v=1}^{V}\phi_{k,v}^{\beta_v-1})d\phi \qquad (2)$$
$$=\int\prod_{k=1}^{K}\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\prod_{v=1}^{V}\phi_{k,v}^{n_{k,v}+\beta_v-1}d\phi$$
$$=\prod_{k=1}^{K}\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma(\beta_v+n_{k,v}^{(vi)})}{\Gamma(\sum_{v=1}^{V}\beta_v+n_{k,v}^{(vi)})}d\phi$$

$$\int p(z,\theta\mid\alpha)d\theta$$
$$=\int p(z\mid\theta)p(\theta\mid\alpha)d\theta$$
$$=\int\prod_{m=1}^{M}(\prod_{k=1|m,k}^{K}\phi^{n_{m,k}})\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}(\prod_{k=1}^{K}\phi_{m,k}^{\alpha_k-1})d\theta \qquad (3)$$
$$=\prod_{m=1}^{M}\int\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{v=1}^{V}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{m,k}^{n_{m,k}+\alpha_k-1}d\theta$$
$$=\prod_{m=1}^{M}\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{v=1}^{V}\Gamma(\alpha_k+n_{m,k}^{(ki)})}{\Gamma(\sum_{k=1}^{K}\alpha_k+n_{m,k}^{(ki)})}d\phi$$

$$\int p(t,\psi\mid z,\gamma)d\psi$$
$$=\int p(t\mid z,\gamma,\psi)\,p(\psi\mid z,\gamma)d\psi$$
$$=\int\prod_{k=1}^{K}(\prod_{y=1|k,y}^{Y}\psi^{n_{k,y}})\frac{\Gamma(\sum_{y=1}^{Y}\gamma_y)}{\prod_{y=1}^{Y}\Gamma(\gamma_y)}(\prod_{y=1}^{Y}\psi_{k,y}^{\gamma_y-1})d\psi \qquad (4)$$
$$=\prod_{k=1}^{K}\frac{\Gamma(\sum_{y=1}^{Y}\gamma_y)}{\prod_{y=1}^{Y}\Gamma(\gamma_y)}\frac{\prod_{y=1}^{Y}\Gamma(\gamma_y+n_{k,y}^{(yi)})}{\Gamma(\sum_{y=1}^{Y}\gamma_y+n_{k,y}^{(yi)})}d\psi$$

Following the detailed derivation process, (2), (3) and (4), we can obtain the remaining probability $\int p(w,\text{Ø}|z,\beta)\,d\text{Ø}$, $\int p(z,\theta|\alpha)\,d\theta$ and $\int p(t,\Psi|z,\text{٢})\,d\Psi$, respectively.

In Gibbs sampling a MCMC is organized to have a particular stationary distribution. In this paper, we want to acquire a MCMC that can converge to the posterior distribution over zgiven M^{train}, α, β, ٢.

Obtaining a sampling from the distribution $p(z|M^{train}, \alpha, \beta, \text{٢}, \sigma, \eta)$, where ٢,σare constant, using Gibbs sampling by (a) sampling a topic distribution zi for an individual word wi, given on constant distributions of topics for all other words in the data set, and (b)iterating this procedure for every word. In equation (5) we show how to derive the basic equation though equations (1), (2), (3), (4) needed for the Gibbs sampler. Where w_i=v represents the observation that ith word is the vth word in the lexicon, z_i=k represents the assignment of the ith word within a document to a topic k, t_i=y represents yth year of document publishing, attached with the ith word in the document and z_{-i} represents all topic assignment not including the ith word.

In above formula the denominator minus one expresses deleting the topic distribution for present word, and deleting the topic distribution, corresponding to the present word, for

present document. $n_{mk}^{(ki)}$ is the total number of topics related with document m. $n_{kv}^{(vi)}$ is the total number of words related with topic k. $n_{ky}^{(yi)}$ is the total number of time slices related with topic k.

M-step:

$$
\begin{aligned}
& p(z_i = k \mid w_i = v, z_{-i}, t_i = y, \alpha, \beta, \eta, \sigma^2, \gamma) \\
& \propto \frac{p(w, z, t \mid \alpha, \beta, \eta, \sigma^2, \gamma)}{p(z_{-i}, w_{-i}, t_{-i} \mid \alpha, \beta, \eta, \sigma^2, \gamma)} \\
& = \frac{\alpha_k + n_{m,k}^{(ki)}}{\sum_{k=1}^{K}(\alpha_k + n_{m,k}^{(ki)}) - 1} \frac{\beta_v + n_{k,v}^{(vi)}}{\sum_{v=1}^{V}(\beta_v + n_{k,v}^{(vi)}) - 1} \frac{\gamma_y + n_{k,y}^{(yi)}}{\sum_{y=1}^{Y}(\gamma_y + n_{k,y}^{(yi)}) - 1} \\
& \exp(-\frac{(L_m - \eta^T \bar{\bar{z}}_m)^2}{2\sigma^2} + \frac{(L_m - \eta^T \bar{\bar{z}}_{m,-i})^2}{2\sigma^2}) \\
& = \frac{\alpha_k + n_{m,k}^{(ki)}}{\sum_{k=1}^{K}(\alpha_k + n_{m,k}^{(ki)}) - 1} \frac{\beta_v + n_{k,v}^{(vi)}}{\sum_{v=1}^{V}(\beta_v + n_{k,v}^{(vi)}) - 1} \frac{\gamma_y + n_{k,y}^{(yi)}}{\sum_{y=1}^{Y}(\gamma_y + n_{k,y}^{(yi)}) - 1} \\
& \exp(\frac{\eta_k}{N_m \sigma^2}(l_m + \sum_{j=1}^{K} \frac{\eta_j}{N_m} n_{j,k}^{(ki)} - \frac{1}{2}\frac{\eta_k}{N_m}))
\end{aligned}
\tag{5}
$$

When parameters α, β and r are fixed according to the decomposed formula p(z,w,l|α, β,ϒ,σ^2,η), the Probability distributions of ∫p(w,Ø|z,β)dØ, ∫p(z,θ|α)dθ and ∫p(t,Ψ|z,ϒ)dΨ are constant. So, the maximize the(marginal) log likelihood of p(z,w,l|α, β,ϒ,σ^2,η) is simplified as max log p(l|z,σ^2,η). By using the normal distribution the further expansion of this formula is showed as following (6).

$$
\begin{aligned}
& \max_{\eta,\sigma^2} \log p(L \mid z, \eta, \sigma^2) \\
& = \max_{\eta,\sigma^2} \log \left[\prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2}\,\pi} \exp\left(-\frac{(L_m - \eta^T \bar{\bar{z}}_m)^2}{2\sigma^2} \right) \right] \\
& = \max_{\eta,\sigma^2} \left(-\frac{D}{2}\log(\sigma^2) - \frac{\left\| l - \bar{\bar{z}}^T \eta \right\|_2^2}{2\sigma^2} \right)
\end{aligned}
\tag{6}
$$

Here the expectation is over the matrix $\bar{z}$, using linearity of expectation and applying the first-order condition for η, we arrive at an expected-value according to $l=\bar{z}^T\eta$.

$$\eta = \left(\bar{\bar{z}}\,\bar{\bar{z}}^T \right)^{-1} \bar{\bar{z}}\, l$$

Furthermore, we can obtain the value of σ^2 by applying the first-order condition for σ^2 to (6) and evaluate the solution at η, obtaining:

$$\sigma^2 = \frac{1}{D}\left(ll^T - l^T \bar{\bar{z}}^T \left(\bar{\bar{z}}\,\bar{\bar{z}}^T \right)^{-1} \bar{\bar{z}}\, l \right)$$

4. Response Variable and Attention of Topic

Considering the topic is the distribution of words, the determination of response variable for topic can be done by the response variable of words associated for topic.

But every word has different weight for response variable, according the discrimination of the word to document in *tfidf*, we take into account *tfidf* during calculating the discrimination of word to response variable, combine the word frequency to obtain the relationship matrix of word to response variable as following.

$$WL_{v,c} = \overline{\sum_{d \in C} \sum_{v \in d} n_v \times tfidf_v} \qquad (7)$$

C denotes the response variable of the document; v is the word in document d. For every word with this response variable we obtain the weight of word to response variable by the product of word frequency and *tfidf*, and take the average.

Considering the topic is the distribution of words, the response variable of topic can be calculated by the response variable of words. So the topic's response variable can been obtained as follow.

$$CL = \sum_{v} \phi_v \times WL_v \qquad (8)$$

As a document d is related with a multinomial distribution θ over topics, we can further predict a document's response variable by its max topics' response variable max (θ_d).

Because our analysis reduces the document to a distribution of topics, it is straightforward to analyze the attention degree of these topics as a means of gaining insight into the trends of the documents. we can formulate more sophisticated generative models that incorporate parameter describing the attention degree of topic based on the estimation of produced θ .The attention degree of topic is defined as the probability proportion of topic relative to corpus. We can gain the attention of topic as following.

$$TH_k = \frac{\sum_d \theta_{d,k}}{D} \qquad (9)$$

D is the accumulated value of all $\theta_{d,k}$.

5. Experiments

In this paper we use a data set from Xinhua News Agency for evaluation. An insufficient expression of the data sets adopted in our experiments is presented as following.

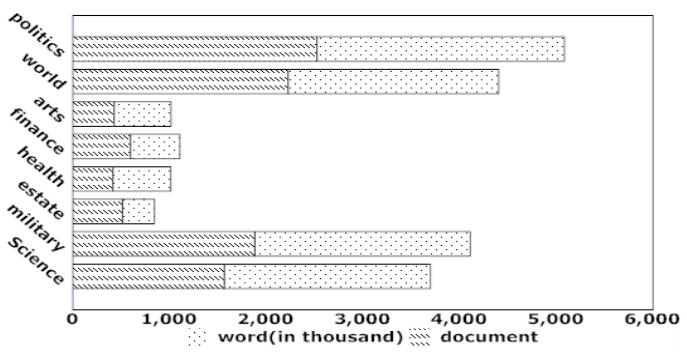

Figure 2. The Distribution of Document and Word

Xinhua News corpus includes of news articles and ordered by time. For our experiments, we use the data set contains articles from Xinhua News between 2013, 6, 1 to 2013, 11, 30. In order to avoid the influence of response variable, in this case refers to classification variable of article, including little papers to our model, we only extract eight classifications consist of arts, business, health and so on. The data was preprocessed for down-casing, deleting extremely common words and numbers, and removing the words that frequency less than six times in the data set. After this preprocess the data set contains 9,805 articles, 48,765 unique words and total of 1,989,735 words. Each document associated with classification label has a time stamp that is determined by the day. Fig.2 shows data and document distribution for corpus.

5.1. Comparing predictive power for Different Models

The density measurement, expressing the potential configuration of data, is the intention of document modeling. Measuring the model's universal performance on formerly unobserved document is general method to estimate this. Perplexity is a canonical measure of goodness that is used in language modeling to measure the likelihood of a held-out test data to be generated from the potential distributions of the model [4]. Better universal performance is manifested by a lower perplexity and the higher the likelihood on a held-out document. Formally, for a test set of M documents, the perplexity is [3]:

$$perplexity(D_{text}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\} \qquad (10)$$

The foremost in formula (10) is how to calculate the likelihood p (w_d) of test document. Based on distribution Ψ and Ø, that are fit using the E step in formula (5), we can employ Gibbs Sampling to obtain θ_{td}, the topic distribution of test documents, by following:

$$p(z_i = k \mid w_i = v, z_{-i}, t_i = y, \alpha, \beta, \eta, \sigma^2, \gamma)$$
$$= \frac{\alpha_k + n_{m,k}^{(ki)}}{\sum_{k=1}^{K}(\alpha_k + n_{m,k}^{(ki)}) - 1}\phi_{k,v}\psi_{k,y}$$
$$\exp\left(\frac{\eta_k}{N_m\sigma^2}\left(l_m + \sum_{j=1}^{K}\frac{\eta_j}{N_m}n_{j,k}^{(ki)} - \frac{1}{2}\frac{\eta_k}{N_m}\right)\right)$$

So, the derivation of the probability p (w_d) can been calculated by multiplying θ_{td}, Ψ and Ø for every word in test set.

In our experiments, the hyper-parameter α, β, ٢were set at 0.45, 0.04 and 0.1. In the first

$$p(w_{td}) = \sum_{\substack{y \in td \\ y \in td}} \theta_{td} \times \phi_v \times \psi_y$$

group of experiments we compared the Latent Dirichlet Allocation and Topics over Time of Section 2.1, supervised Latent Dirichlet Allocation proposed by Blei and McAuliffe in 2007, and our presented Time Label Topic model.

Figure 3 shows the experiment result of the perplexity as a function of the number of training data corpus for the compared 4 models. Since the time stamp and response variable are not been considered in LDA model, which limits its generalization performance, the LDA is distinctly not better than either of topic-based models, as demonstrated by its high perplexity. TOT model optimizes the generalization performance of LDA by adopting time stamp information. TLT model uses knowledge of the response variable associated with document to extend a better prior for the data set based on TOT. So TLT further develops the generalization performance of TOT model to document associated with response variable. As M^{trian} increases we see the same result.

5.2 Time Stamp and Response Variable Prediction

TLT model can been further used for time stamp and response variable prediction that is classification in this paper, dating a article or detected topics given the words.

The methods of classification prediction for discovered topics and article have introduced in Section 4. Shown a article, the time stamp can be predicted by discretizing time stamp, and maximizing the posterior that is computed by $\max_t \prod_{i=1}^{Nd} p(t \mid \psi_{zi})$. The used parameters Ø and Ψ can been calculated by Equation (5) in M^{train} corpus.

This work offers a different method to quantitatively compare TLT model and other model mentioned above. We predict the month stamp and classification for testing article and estimate the error in terms of the accuracy, L1 error (the difference between predicted and true month or true classification) and average L1 distance over the testing data set to the right month, and right classification respectively.

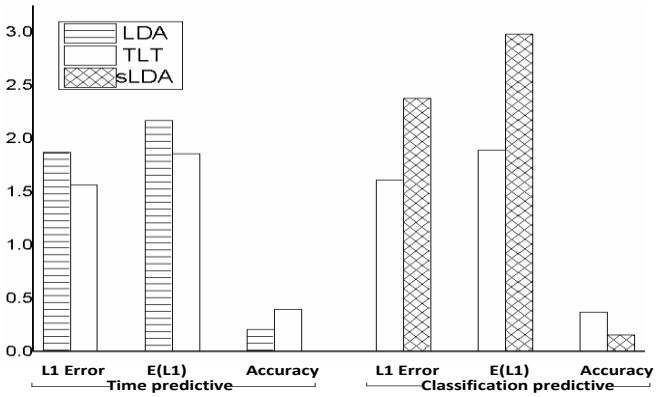

Figure 5. The Prediction of Time and Classification

Figure 5 illustrates the results of time predictive and classification predictive for testing set. As shown in it, the accuracy of TLT model always performs better than or as well as anyone model else and affords L1 error 22% decrease with respect to TOT for time predictive. The prediction of classification shows similar experiments with time predictive.

5.3. Topic Revolution over Time

Primely modeling can well discover topics with classification from the XinHua data set, and facilitate us comprehend how topic evolve, and assess the intensity of variously shaped profiles over time. This benefits the reader grasp more accurately when and over what length of time the topic trend was proceeding.

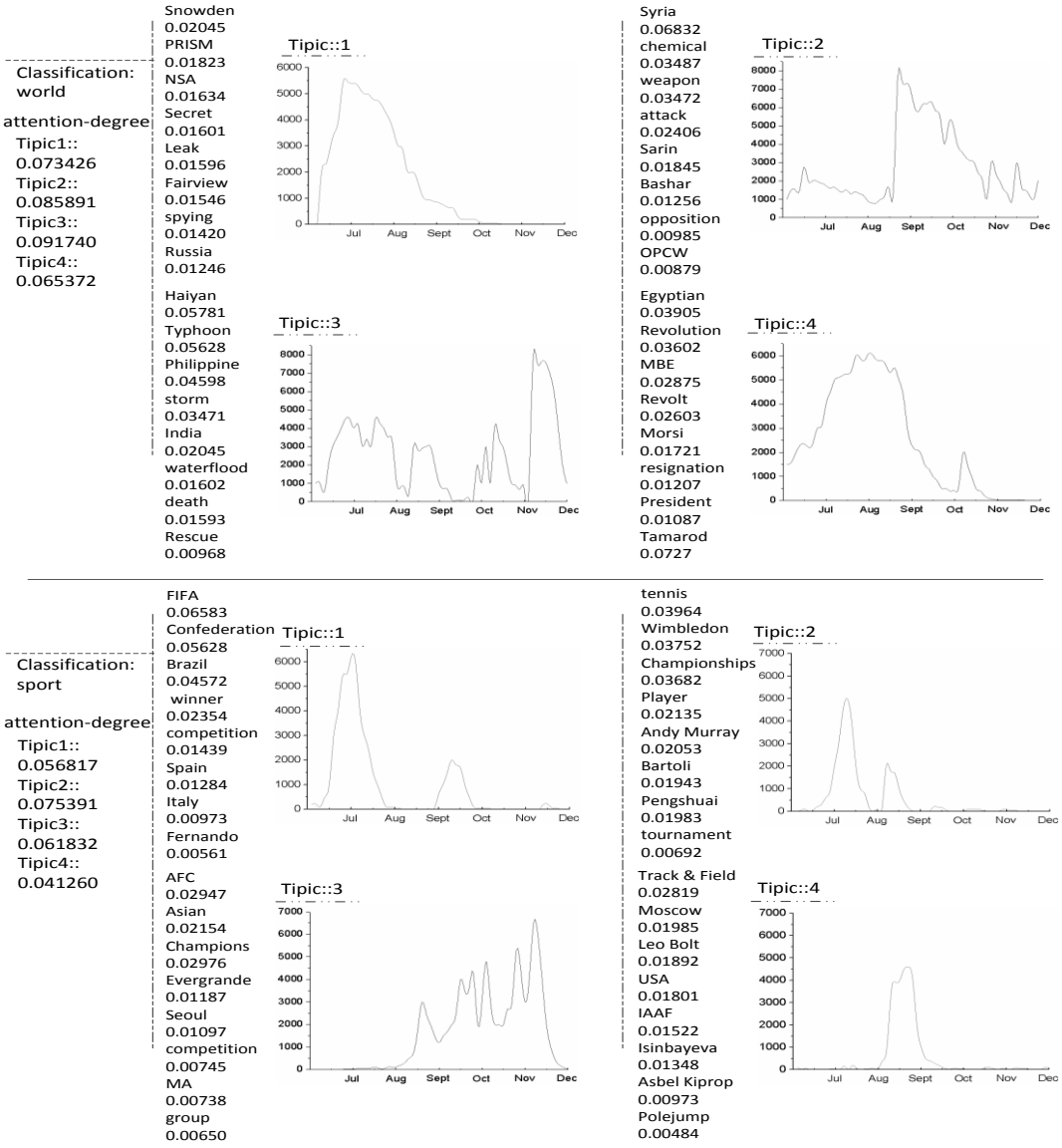

Figure 4. Evaluation of Topic Over Time for Different Classifications

Two major classifications, four most likely discovered topics associated every classification, the top eight their words most likely to be generated, their actual line graphs of

intensity over time on Xinhua data set are illustrated in Figure 4. The attention-degree of topic is calculated by Equation (9). The distribution of topic over time is in accord with Multinomial distribution decided by Ψ.

The first four topics with high attention in the world classification are selected. The first topic is the Snowden Leaks event occurred in the few days after June and popular over the time-period June-October. The second topic shows the rise and fall of Syria chemical weapons event that happens around the end of august. However Syria war confuses this topic distribution. The third topic is Haiyan Typhoon topic. But during June to December in 2013 there are many flood and typhoon event, especial rainstorm at North India around the end of June, these flood event make the distribution not clear. The fourth topic is Egyptian protest topic that came to a head In mid August.

In classification of sport the first four topics with high attention are FIFA Confederations Cup, Wimbledon Open, AFC Champions league and Moscow's Track and Field. They show the similar results as classification of world.

Summary

In this paper, the TLT model proposed offers a relatively simple probabilistic model for exploring the relationships between documents, topics, words, time and response variable label. This model deals with the problem of discovering the attention degree of constructed topics over time, and provides expressively optimized performance in the field of perplexity, time and response variable prediction compared to LDA and sLDA models. Empirical results and discussions prove the capacity of proposed model. From an ordinary perspective, our model can also be applied to other dataset except Xinhua news. However, during practical testing, we do not test TLT model on other dataset. In addition, this paper only incorporates time and response variable information. So possible future directions for this work include test model on other dataset, integrates other side information, such as user's label, to perform detection and analysis of analysis and detect topics.

Acknowledgments

This work is supported by science and technology project of Fujian provincial education department: No. JA13196; the National Natural Science Foundation of China 61303131; the National Natural Science Foundation of China under Grant No. 60373099, No. 60973040.

References

[1] T. H. Probabilistic latent semantic indexing, In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SI-GIR'99), **(1999)**, pp. 50-57.
[2] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, vol. 3, **(2003)**, pp. 993-1022.
[3] M. S. and T. L. G. Probabilistic Topic Models", Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum, **(2005)**.
[4] G. H. Parameter estimation for text analysis, Arbylon publications, **(2007)**.
[5] X. W., A. M. 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, **(2006)**, pp. 424–433.
[6] D. M. Blei and J. M. Supervised Topic Models. In NIPS, vol. 21, **(2007)**.
[7] D. M and A. M. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression, Uncertainty in Artificial Intelligence, **(2008)**, pp. 411-418.
[8] X. R. W. Structured Topic Models: Jointly Modeling Words and Their Accompanying Modalities, University of Massachusetts Amherst, Computer Science, **(2009)**.

[9] L. A., D. B. On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. In ICDM Data Mining, **(2008)**.
[10] D. M. Blei, Probabilistic Topic Models", Communications of the ACM, vol. 55, no. 4, **(2012)**, pp. 77-84.
[11] Y. Zhou, S. Ji and K. Xu, "Massive Scientific Paper Mining: Modeling", Design and Implementation, Lecture Notes in Computer Science, vol. 82, **(2013)**, pp. 343-352.
[12] P. F. Hu and W. Liu, "Latent topic model for audio retrieval Pattern Recognition", vol. 3, no. 47, **(2014)**.