Trend Detection Model

Noriaki Kawamae*
NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0237
kawamae@gmail.com

Ryuichiro Higashinaka†
NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0237
higashinaka.ryuichiro@lab.ntt.co.jp

ABSTRACT

This paper presents a topic model that detects topic distributions over time. Our proposed model, Trend Detection Model (TDM) introduces a latent trend class variable into each document. The trend class has a probability distribution over topics and a continuous distribution over time. Experiments using our data set show that TDM is useful as a generative model in the analysis of the evolution of trends.

Categories and Subject Descriptors

G.3 [**PROBABILITY AND STATISTICS**]: Time series analysis

General Terms

Algorithms, experimentation

Keywords

Topic Model, Trend Model, Dynamic Topic Model, Latent Variable Modeling, Timestamps, Trend Analysis

1. INTRODUCTION

Modeling the evolution of trends over time is important in the analysis of user behavioral data and large document collections such as mails, news, and blogs.

This paper presents Trend Detection Model(TDM); it models trends over continuous time. In this model, we suppose that each trend can be presented as a set of (1) the mixture of topics and (2) localization over time. Following this assumption, we introduce a latent variable, called trend class, that has both a probability distribution over both topics and a beta distribution over time, into each document.

A key advantage of TDM is that it can capture trends of different spans at the same time in the low-dimensionality set of topics and timestamps. TDM captures the topic evolution over time by the trend class, while Dynamic Topic Models (DTMs) [1] captures the word evolution of each topic over time. Simultaneously, this class predicts absolute time values given an unstamped document, and predicts topic distributions given the words in document as Topics Over Time (TOT) [2] can, since this class is associated with a continuous distribution over time. Consequently, this model incorporates two characteristics for modeling trends in documents and realizes the functionalities of both DTMs and TOT at the same time.

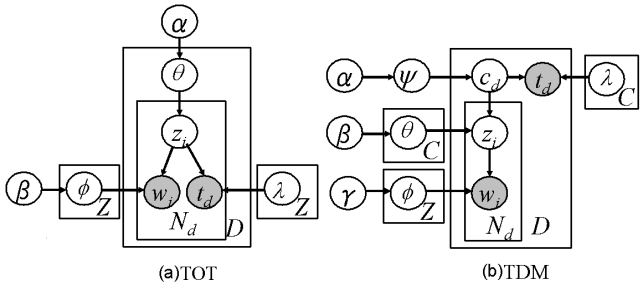

Figure 1: Graphical Models: TOT and TDM: In this figure, shaded and unshaded variables indicate observed and latent variables, respectively. An arrow indicates a conditional dependency between variables and stacked panes indicate a repeated sampling with the iteration number shown.

2. TREND DETECTION MODEL

2.1 The trend class

In this subsection, we describe our model. Table 1 shows the notations used in this paper; Figure 1 shows the generative process using graphical models of (a) TOT and (b) TDM. Before introducing our model, let us review the concept of Topics Over Time (TOT) model. TOT explicitly models absolute timestamp values by parameterizing the continuous distribution over time associated with each topic. In this model, topics are responsible for generating both observed timestamps as well as words.

Our proposed model, basic TDM models time jointly with topic co-occurrence patterns rather than word co-occurrence patterns. A novel feature of this model is the inclusion of trend class c; it is responsible for generating both observed timestamps and topics in each document. The trend class allows TDM to represent trends by topic distribution associated with time θ_{TDM} rather than word distribution associated with time ϕ_{TOT}. Therefore, TDM assigns the same trend class to documents if they have almost identical timestamps as well as a similar set of topics, otherwise they must

*Currently with NTT Comware Corporation.
†Currently with NTT Cyber Space Laboratories, NTT Corporation.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

Table 1: Notation used in this paper

SYMBOL	DESCRIPTION
C	number of trend classes
Z	number of topics
D	number of documents
V	number of unique words
N_d	number of tokens in d
t_d	timestamps associated with d
c_d	trend associated with d
z_{di}	topic associated with the ith token in d
w_{di}	ith token in document d
ψ	multinomial distribution of trend classes ($\psi\|\alpha \sim$ Dirichlet(α))
$\lambda_{z,c}$	beta distribution associated with z, c
θ_c	multinomial distribution of topics specific to c ($\theta_c\|\beta \sim$ Dirichlet(β))
ϕ_z	multinomial distribution of words specific to z($\phi_z\|\gamma \sim$ Dirichlet(γ))

be assigned to different trends.

2.2 Inference and Learning

The generative model for TDM can be described by the Bayesian hierarchical model. We employ Gibbs sampling to perform approximate inference in TDM. In the Gibbs sampling procedure, we need to calculate the conditional distributions. We use the chain rule and can then obtain the conditional distribution $P(c_d = j|\mathbf{c}_{\setminus d}, \mathbf{z}, \mathbf{t}, \alpha, \beta, \lambda)$ as

$$P(j|\cdots) \propto \frac{n_{j\setminus d}+\alpha_j}{\sum_c^C (n_{c\setminus d}+\alpha_c)} \frac{\Gamma(\sum_z^Z n_{jz\setminus d}+\beta_z)}{\prod_z^Z \Gamma(n_{jz\setminus d}+\beta_z)} \frac{\prod_z^Z \Gamma(n_{jz}+\beta_z)}{\Gamma(\sum_z^Z n_{jz}+\beta_z)} \times \frac{(1-t_d)^{\lambda_{j1}-1} t_d^{\lambda_{j2}-1}}{B(\lambda_{j1},\lambda_{j2})}$$

where $n_{j\setminus d}$ represents the number of documents (except for d) that have been assigned to j, $n_{jz\setminus d}$ represents the number of tokens assigned to topic z in the documents (except for d) associated with j, and B is the beta function.

Likewise, the predictive distribution of adding word w_{di} to topic k is $P(z_{di} = k|j, \mathbf{z}_{\setminus di}, \mathbf{w}, \beta, \gamma)$ and is written as

$$P(k|\cdots) \propto \frac{n_{kw_{di}\setminus di}+\gamma_{w_{di}}}{\sum_v (n_{kv\setminus di}+\gamma_v)} \frac{n_{jk\setminus di}+\beta_k}{\sum_z (n_{jz\setminus di}+\beta_z)}, \quad (2)$$

where $n_{kv\setminus di}$ represents the number of tokens assigned to word v in topic k, except d_i.

3. EXPERIMENTS

We present the quantitative evaluations of the proposed models, where we used a data set: 8 years (2001-2008) of research papers in the proceedings of ACM CIKM, SIGIR, KDD, and WWW. After removing stop words, numbers, and the words that appeared less than five times in the corpus from this data, we yielded a total set of 3078 documents and 20286 unique words from 2204 authors. In our evaluation, the smoothing parameters α,β, and γ were set to {1/Z(TOT), 1/C(TDM)}, {0.1(TOT), 1/Z(TDM)}, and 0.1, respectively.

Table 2: Perplexity comparison: All models are learned with the number of topics Z set at 200. The number in the second row for TDM is the number of trend classes. These results are averaged over five-fold cross validation. Results that differ significantly by parametric non-paired t-test $p < 0.01$ from TOT are marked with '*'.

DTMs	TOT	TDM			
		25	50	75	100
1587	1543	1488*	1457*	1441*	1436*

To measure the ability of the proposed model to act as generative model, we computed test set perplexity under the estimated parameters and compared the resulting values, and show the results in Table 2. From this table, we observe that the trend class allows TDM to group documents under the various topic distributions rather than permitting various topic distributions on each document. This implies that clustered documents contain less noise than otherwise, and reduce the perplexity over all.

One interesting common feature of both TOT and TDM is the ability to predict the timestamp given the words in a document. This functionality also provides another opportunity to quantitatively compare TDM against TOT. On the corpus, we measure the ability to predict the published year given paper and show the results in Table 3.

Table 3: L1 error comparison: The number in the first row is the number of topics. The number in parentheses for TDM is the number of trend classes. Results that differ significantly by parametric non-paired t-test $p < 0.01$, $p < 0.05$ from other methods are marked with '' and '*' respectively.**

(1)

Model	50	100	150	200
TOT	2.44	2.25	2.11	1.97
TDM(50)	2.11*	1.93**	1.88**	1.76**
TDM(100)	2.03**	1.85**	1.71**	1.65**

Since TOT would generate different time stamps within the same document, this generation is overwhelmed by the plurality of words generated under the bag of words assumption. This defect dampens the predictive performance.

From these results, we can say that TDM attains lower perplexity than DTMs and TOT, and can predict what topic will rise more accurately.

4. CONCLUSION

In this paper, we proposed a model that takes time jointly with topic co-occurrence. Experiments using various data sets showed that TDM captures the trends of different spans at the same time. In future work, we will extend TDM by considering other metadata.

5. REFERENCES

[1] D. Blei and J. Lafferty. Dynamic topic models. In *ICML*, pages 113–120, 2006.

[2] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *KDD*, pages 424–433, 2006.