Trend Analysis Model: Trend Consists of Temporal Words, Topics, and Timestamps

Noriaki Kawamae
Tokyo Denki University
2-2 Kanda-Nishiki-cho, Chiyoda-ku, Tokyo, Japan
Japan 101-8457
kawamae@gmail.com

ABSTRACT

This paper presents a topic model that identifies interpretable low dimensional components in time-stamped data for capturing the evolution of trends. Unlike other models for time-stamped data, our proposal, the trend analysis model (TAM), focuses on the difference between temporal words and other words in each document to detect topic evolution over time. TAM introduces a latent trend class variable into each document and a latent switch variable into each token for handling these differences. The trend class has a probability distribution over temporal words, topics, and a continuous distribution over time, where each topic is responsible for generating words. The latter class uses a document specific probabilistic distribution to judge which variable each word comes from for generating words in each token. Accordingly, TAM can explain which topic co-occurrence pattern will appear at any given time, and represents documents of similar content and timestamp as sharing the same trend class. This class allows TAM to project these data on a latent space of trend dimensionality and predict the temporal evolution of words and topics in them. Experiments on various data sets show that the proposed model can capture interpretable low dimensionality sets of topics and timestamps, take advantage of previous models, and 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 Modeling, Latent Variable Modeling, Trend Analysis, Bayesian hierarchical model, Graphical Models, Timestamped data, Timestamp

1. INTRODUCTION

Modeling the evolution of trends over time is critical for the analysis of user behavioral data and large document collections such as mails, news, and blogs. Since users' interests and their subjects change over time, many services (e.g., Recommendation, Search, Advertise and Marketing) try to analyze trends and use the results for improving service quality. Therefore, they need a model that can automatically track trends and predict them. As one example of the application of this model, an information retrieval service would show hot topics or recommend web content to satisfy users' current preferences.

The goal of modeling trends in documents is to detect the rise and fall of each trend and which topics each trend consists of. In general, each document consists of several topics and the mixture of these topics change over time even among the documents associated with the same theme. For example, in the theme of "economic news", the subprime loan problem and the bankruptcy of Lehman Brothers are temporal topics that have dramatically emerged and then disappeared over time, while the topics of exchange rates and interest endure over time. Moreover, the change in what is popular in the economic news is faster than that in arts. Therefore, a trend has the following characteristics: 1) Among documents about the same theme, the proportion of topics changes over time. 2)The rate of this change varies with the theme.

Although we need to take these two characteristics into account in modeling trends in documents, previous algorithms have incorporated only one or the other of these characteristics. For example, Dynamic Topic Model, (DTM) [1], explicitly models the evolution of topics over time by estimating the topic distribution at various epochs, where each topic is a multinomial distribution over a word vocabulary. This quantization of time always poses the issue of how to select the slice size; the appropriate size depends on the documents. Accordingly, this model incorporates the evolution of each topic over time, but fails to simultaneously model trends with different spans. Another proposal, Topics Over Time (TOT) [18] associates each topic with a beta distribution over time; this allows the co-occurrence probability of a topic at any given time to be represented. Since this model learns the parameters of this distribution on each topic based by using the timestamps of corresponding document in each token, it fails to identify the change in topic distribution over time. Therefore, it is necessary to incorporate both of these characteristics in a unified model.

WSDM'11, February 9–12, 2011, Hong Kong, China.

This paper presents Trend Analysis Model(TAM); it models trends over continuous time by focusing on the difference between temporal words and other words, and can detect topic co-occurrence distributions over time. In this model, we assume that each trend can be presented as a mixture of temporal words, terminology words, and localization over time. For example, "Twitter" and "Social network" are temporal words found in the latest data mining conferences, while "Clustering" and "Ranking" are terminology words used over the years. Simultaneously, temporal words such as "Wiki" are much more likely to have appeared in the latest papers, whereas words such as "abstract" and "introduction" appear constantly in papers over long time periods. Following this assumption, we introduce a latent variable, called trend class, which has both a probability distribution over both temporal words and topics, and a beta distribution over time, into each document. Since the beta distribution can take versatile shapes, we use it to describe the trend-class-specific continuous time in each document. Additionally, we add another latent variable, called switch variable, which represents a probability distribution over the topic, the trend class, and a background class. The switch variable handles variables(background, trend class and topic) for generating words in each token, and distinguishes words associated with the trend class as temporal words. Therefore, TAM puts the documents that have similar topic distributions over time into the same trend class.

A key advantage of TAM is that it can capture trends with different spans at the same time in the low-dimensionality set of words, topics, and timestamps. (1)TAM uses trend class to capture topic evolution over time, while DTMs capture the evolution of the representative words of each topic over time. (2)Simultaneously, the trend class predicts absolute time values to given an unstamped document, and predicts topic distributions to given the words in document as can TOT, since this class is associated with a continuous distribution over time. (3)Moreover, since the switch variable is trained to identify background/trend/topic specific words, TAM allows us to infer more distinct topics than existing schemes that ignore this variable. Consequently, TAM offers two characteristics for modeling trends in documents and realizes the functionalities of both DTMs and TOT at the same time. We demonstrate the efficacy of TAM through experiments and show that this model can describe the structure of a wide variety of trends.

2. RELATED WORK

Earlier work examined topics and their changes across time. Although time is intrinsically continuous, most studies modeled the evolution of topics over time by estimating the topic distribution at various epochs. For example, Dynamic topic Models (DTMs) [2] estimates the topic distribution at various epochs, where the natural parameters of this multinomial distribution are conditionally distributed by a normal distribution with mean equal to the natural parameters at the previous epoch. Multi scale Topic Tomography Model (MTTM) [15] employs non-homogeneous Poisson processes to model the generation of word-counts, combined with multi-scale analysis using Haar wavelets.

Recent work on topic models has investigated richer structures to describe inter-topic correlations. An example is hierarchical LDA (hLDA) [3], which assumes a hierarchical structure among topics, where topics at higher levels are more general, such as stop words and topics at lower levels are organized into topics, such as more specific words. The pachinko allocation model (PAM) [13] uses a directed acyclic graph structure to represent topic correlations. Since these models and TAM are complementary, hierarchical TAM is a natural extension.

It seems likely that distinguishing stop words and incorporating vocabulary evolution are useful for improving the quality of topic models, rather than assuming all words are terminology words and then assigning topics in each document. Kleinberg [10] developed an approach for modeling bursts by using an infinite-state automation, where bursts appear naturally as state transitions. Although contextual probabilistic latent semantic analysis (CPLSA) [14] discretizes time, this model proposes a background model that captures words that are present all the time by extending the probabilistic latent semantic analysis. Topic monitor [4] proposes folding-in techniques for topic adaptation under an evolving vocabulary.

Recently, there has been a focus on mining and summarizing sentiment in Weblogs and user reviews [8], or systems that produce fine-grained sentiment analysis of user reviews [19], where the goal is mainly to mine user opinions by identifying and extracting positive and negative opinions or analyzing and extracting topical contents. Several topic models have been proposed for this goal [5, 16]. Although many of these data are collected over time and are indeed dynamic, these works do not incorporate this dynamic into extracting the ratable aspects of objects from these review data. Here, TAM can play a complementary role to these models and we can create multi-grain TAM. Collaborative filtering [11] is also one of the most promising applications and focuses on the introduction of the time factor [7, 12]. Therefore, these approaches and TAM are complementary in this regard.

3. TREND ANALYSIS MODEL (TAM)

3.1 Modeling the trend by using the trend class and switching value

In this subsection, we describe our model TAM; it simultaneously achieves dimensionality reduction in the representation of a multinomial distribution over temporal words, topics and the continuous distribution over time associated with each mixture of topics. Table 1 shows the notations used in this paper; Figure 1 shows the graphical models of previous model Topics over Time (TOT) and our models: switch Topics Over Time (sTOM), basic Trend Detection Model (TDM) and Trend Analysis Model (TAM).

Before introducing our model, let us review the concept of Topics Over Time (TOT) model. TOT explicitly models time jointly with word co-occurrence patterns. Rather than discretizing time and modeling a sequence of state changes with a Markov assumption on the dynamics, TOT 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.

Switch TOT (sTOT) is an extended model of TOT, by introducing latent switch variable r. In this model, we focus on differencing a trend specific words and a background topic words in each generative process of timestamped data. The background topic is the common topic over almost all docu-

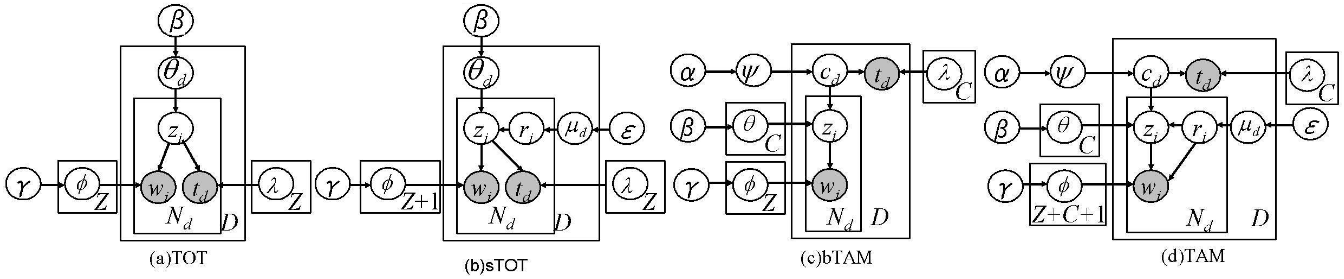


Figure 1: Graphical Models: TOT, sTOT, TDM and TAM: In this figure, shaded and unshaded variables indicate observed and latent variables, respectively. An arrow indicates a conditional dependency between variables and stacked panes indicate a repeated sampling with the iteration number shown.

Table 1: Notation used in this paper

SYMBOL	DESCRIPTION
C	number of trend classes
D	number of documents
Z	number of topics
V	number of unique words
N_d	number of word tokens in document d
t_d	the timestamps associated with document d
c_d	the trend associated with document d
r_i	the switch associated with the ith token
z_i	the topic associated with the ith token
w_i	the ith token
ψ	the multinomial distribution of trend classes ($\psi\|\alpha \sim$ Dirichlet(α))
λ_c	the trend c specific beta distribution of t
μ_d	the multinomial distribution of r associated with document d ($\mu_d\|\epsilon \sim$ Dirichlet(ϵ))
$\theta_{c(d)}$	the multinomial distribution of topics specific to trend c(document d) ($\theta_{c(d)}\|\beta \sim$ Dirichlet(β))
$\phi_{z(c,b)}$	the multinomial distribution of words specific to topic z(trend class c,background topic b) ($\phi_{z(c,b)}\|\gamma \sim$ Dirichlet(γ))
$\alpha, \beta, \gamma, \epsilon$	the fixed parameters of symmetric Dirichlet priors

ments regardless of their content and time, it generates non temporal words, while the topic variable generates temporal words. For distinguishing temporal words from background topic words, we define r, it acts as a switch to handle these words in each token, and takes value r=0 if word w is generated via the background topic variable, or r=1 if word w is generated via the topic variable.

An alternative model of Trend Analysis Model, Trend Detection Model (TDM) [9] associates a single timestamp 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 the observed timestamp as well as topics in each document. This class allows TDM to represent trends by topic distribution associated with time θ_{TDM} instead of word distribution associated with time ϕ_{TOT}. Since the Beta distribution can take versatile shapes, TDM uses this to describe time t associated with the trend class in each document too; TOT uses it in each token, where all timestamps are normalized from 0 to 1. Therefore, TDM enables us to explore topic and timestamp co-occurrence patterns within a trend by using the trend class. This model assigns documents to the same trend class if they have almost identical timestamps as well as a similar set of topics, otherwise they must be assigned to different trends.

Here, we propose Trend Analysis Model (TAM), which is generalization of these previous models, by introducing both c and r. The combination of these variables can provide the advantages of each model by describing the generative process more precisely. Although some words appear intensively within a specific period of time and some words appear constantly over long periods in documents with specific theme, TDM treats the former words as words associated with temporal topics. For distinguishing the differences in word tokens, we extend r as a switch for handling more kinds of words as follows. If r=0, TAM selects the background class as responsible for generating co-occurring words. If r=1, TAM selects the topic as responsible for generating terminology words associated with each topic constantly over long periods as is assumed by previous topic models; TAM uses TDM in this token. If r=2, TAM selects the trend class as responsible for generating temporal words in a short time. Note that temporal words are generated directly from c in TAM, while these words are generated from z in TDM. We distinguish trend class from topic or background class for detecting trends in the data, and then employ the hierarchical structure in describing this generative process.

3.2 Inference and Learning

Since TAM is a generalization of TOT, and TDM, we can infer TAM by Gibbs sampling in the same way used for TOT without loss of generalization. The first step, defining the generative process of TAM for parameter estimation, is as follows:

1. Draw multinomial ψ from Dirichlet prior α;
2. Draw C multinomials θ_c from Dirichlet prior β, one for each trend class c;
3. Draw C beta distributions λ_c, one for each trend class c;
4. Draw D multinomials μ_d from Dirichlet prior ϵ, one for each document d;
5. Draw $Z + C + 1$ multinomials $\phi_{z(c,b)}$ from prior γ, one for each topic z (trend class c, background topic b);

6. For each document d:
   - (a) Draw trend class c_d from multinomial ψ;
   - (b) Draw timestamp t_d from beta λ_{c_d};
     For each token i in document d:
     Draw switch r_d from multinomial μ_d;
     if $r_{di} = 1$
       - i. Draw topic z_{di} from multinomial θ_{c_d};
       - ii. Draw word w_{di} from multinomial $\phi_{z_{di}}$.

     else if $r_{di} = 2$
       - i. Draw word w_{di} from multinomial ϕ_{c_d}.

     else
       - i. Draw word w_{di} from multinomial ϕ_b.

The generative model for TAM can be described as a Bayesian hierarchical model. In this inference, we need to calculate the conditional distribution. As shown in the previous subsection, the joint distribution of the entire corpus is, therefore, the following mixture:

$$\begin{aligned}
&p(\mathbf{w}, \mathbf{z}, \mathbf{r}, \mathbf{t}, \mathbf{c}, \phi, \theta, \mu, \psi | \alpha, \beta, \gamma, \lambda, \epsilon)\\
&= p(\mathbf{w}, \phi | \mathbf{z}, \gamma) p(\mathbf{z}, \theta | \mathbf{r}, \mathbf{c}, \beta) p(\mathbf{r}, \mu | \epsilon) p(\mathbf{t} | \mathbf{c}, \lambda) p(\mathbf{c}, \psi | \alpha)\\
&= \prod_d^D [P(c_d | \psi) p(t_d | \lambda_{c_d}) \prod_i^{N_d} [P(w_{di} | \phi_{z_{di}}) P(z_{di} | r_{di}, \theta_{c_d})\\
&\times P(r_{di} | \mu_d)]] \times p(\psi | \alpha) \prod_d^D p(\mu_d | \epsilon) \prod_c^C p(\theta_c | \beta) \prod_z^{Z+C+1} p(\phi_z | \gamma).
\end{aligned} \tag{1}$$

In this eq (1), multinomials ϕ_z, θ_c, μ_d and ψ can be adapted by the conjugate prior and then integrated out analytically as provided in Appendix. In the Gibbs sampling procedure, we need to calculate the conditional distributions $P(c_d | \mathbf{c}_{\backslash d}, \mathbf{z}, \mathbf{t}, \alpha, \beta, \lambda)$ and $P(z_{di}, r_{di} | c_d, \mathbf{z}_{\backslash di}, \mathbf{r}_{\backslash di}, \mathbf{w}, \beta, \gamma, \epsilon)$. In these distributions, $\mathbf{c}_{\backslash d}$ represent the trend class assignments for all tokens except c_d, and $\mathbf{z}_{\backslash di}$ represents the topic assignments for all tokens except z_{di}. Details of the derivation of Gibbs sampling for TAM is given below.

3.2.1 Trend class

For each document, we use the chain rule and then obtain the conditional distribution $P(c_d = j | \mathbf{c}_{\backslash d}, \mathbf{z}, \mathbf{t}, \alpha, \beta, \lambda)$ as

$$\begin{aligned}
P(j | \cdots) &\propto \frac{n_{j\backslash d} + \alpha_j}{\sum_c^C (n_{c\backslash d} + \alpha_c)} \frac{\Gamma(\sum_z^Z n_{jz\backslash d} + \beta_z)}{\prod_z^Z \Gamma(n_{jz\backslash d} + \beta_z)} \frac{\prod_z^Z \Gamma(n_{jz} + \beta_z)}{\Gamma(\sum_z^Z n_{jz} + \beta_z)}\\
&\times \frac{(1 - t_d)^{\lambda_{j1} - 1} t_d^{\lambda_{j2} - 1}}{B(\lambda_{j1}, \lambda_{j2})},
\end{aligned} \tag{2}$$

where $n_{j\backslash d}$ represents the number of documents that have been assigned to j, except d, $n_{jz\backslash d}$ represents the number of tokens assigned to topic z in the documents associated with j, except d, and B is the beta function. Since each topics pattern and its timestamp are assumed to have been generated conditional on trend class c, the resulting beta and multinomial parameters will correspond. A timestamp with high probability under a certain trend class will likely contain the set of topics that co-occurred with high probability in the same trend class. Each new topic/time is generated by again selecting from the trend class c and repeating the entire process. Thus, we can view c as a high level representation of the ensemble of topic/time pairs in terms of a probability distribution over factors that each topic/time can be assembled from.

3.2.2 Switching value and Topic

For each token, the predictive distribution of adding word w_{di} in document d to background topic is $P(r_{di} = 0 | c_d = j, \mathbf{r}_{\backslash di}, \mathbf{z}, \mathbf{w}, \beta, \gamma, \epsilon)$ and is written as

$$P(0 | \cdots) \propto \frac{n_{bv\backslash di} + \gamma_v}{\sum_w^W (n_{bw\backslash di} + \gamma_w)} \frac{n_{d0\backslash di} + \epsilon_0}{\sum_r^R (n_{dr\backslash di} + \epsilon_r)}, \tag{3}$$

where $n_{d0\backslash di}$ represents the number of tokens assigned to switch $r = 0$ (background topic) in document d, except di, and $n_{bv\backslash di}$ represents the number of tokens assigned to word v in background topic, except di. Within the same given corpus, background topic words are generated from the common probability distribution ϕ_b.

Similarly, the predictive distribution of adding word w_{di} in document d to topic k is $P(r_{di} = 1, z_{di} = k | c_d = j, \mathbf{z}_{\backslash di}, \mathbf{r}_{\backslash di}, \mathbf{w}, \beta, \gamma, \epsilon)$ and is written as

$$\begin{aligned}
&P(1, k | \cdots) \propto\\
&\frac{n_{kv\backslash di} + \gamma_v}{\sum_w^W (n_{kw\backslash di} + \gamma_w)} \frac{n_{jk\backslash di} + \beta_k}{\sum_z^Z (n_{jz\backslash di} + \beta_z)} \frac{n_{d1\backslash di} + \epsilon_1}{\sum_r^R (n_{dr\backslash di} + \epsilon_r)},
\end{aligned} \tag{4}$$

where $n_{d1\backslash di}$ represents the number of tokens assigned to switch $r = 1$ (topic) in document d, except d_i, and $n_{kv\backslash di}$ represents the number of tokens assigned to word v in topic k, except d_i. Within a document, the trend class c is fixed; each topic z can vary according to the probability distribution associated with this trend class θ_c. Next, each word can vary according to the probability distribution associated with this topic ϕ_z.

Like eq (4), the predictive distribution of adding word w_{di} in document d to trend class j is $P(r_{di} = 2 | c_d = j, \mathbf{r}_{\backslash di}, \mathbf{z}, \mathbf{w}, \beta, \gamma, \epsilon)$ and is written as

$$P(2 | \cdots) \propto \frac{n_{jv\backslash di} + \gamma_v}{\sum_w^W (n_{jw\backslash di} + \gamma_w)} \frac{n_{d2\backslash di} + \epsilon_2}{\sum_r^R (n_{dr\backslash di} + \epsilon_r)}, \tag{5}$$

where $n_{d2\backslash di}$ represents the number of tokens assigned to switch $r = 2$ (trend class) in document d, except d_i, and $n_{jv\backslash di}$ represents the number of tokens assigned to word v in trend class j, except d_i. Within a document, the trend class c is fixed; each word can vary according to the probability distribution associated with this trend class ϕ_c.

3.3 Trend Transition Discovery

With the trends extracted from all documents, we now turn to the discovery of trend transitions. These transition mean the evolution from trend c_1 to c_2 that has similar topic distribution with c_1, and rises after c_1, that is shown by the beta distribution. To measure the trend transition distance between two trends, we use the Kullback-Leibler divergence. We assume that c_2 has a smaller evolution distance to c_1 if topic distributions θ_{c_1} and θ_{c_2} are closer to each other. Since the KL-divergence $D_{kl}(P||Q)$ can measure the expected number of extra bits required to code samples from P when using a code based on Q, rather than using a code based on P, it appears to be a natural measure of the evolution distance between two trends.

$$D_{KL}(c_1||c_2) = \sum_{i=1}^{Z} p(z_i|c_1) log \frac{p(z_i|c_1)}{p(z_i|c_2)}. \quad (6)$$

Note that the KL-divergence is asymmetric and it makes more sense to use $D_{KL}(c_1||c_2)$ than $D_{KL}(c_2||c_1)$ to measure the evolution distance from c_1 to c_2.

3.4 Alternative of TAM

Although TAM uses the beta distribution to describe the continuous time of each trend, other probabilistic distributions might be preferred if the given timestamped data contains trends exhibiting multimodality. In this data, some topics repeatedly rise and fall in a short period. If this dynamic is to be tracked, the multinomial distribution over quantized periods may a better choice than the beta distribution.

4. EXPERIMENTS

4.1 Data sets

We focus here on the extraction of trends from the given data, and present both qualitative and quantitative evaluations of the proposed models. They were challenged with four data sets:

(1)Data1: 25 years (1978-2002) of research papers in the proceedings of ACM SIGIR. We removed stop words, numbers, and the words that appeared less than five times in the corpus. Accordingly, we obtained a total set of 1658 documents and 12181 unique words. Each document's timestamp was taken to be the year of the proceedings in which it was published.

(2)Data2: 8 years (2001-2008) of research papers in the proceedings of ACM CIKM, SIGIR, KDD, and WWW. The pre processing used for Data1 was applied. This yielded a total set of 3078 documents and 20286 unique words from 2204 authors.

(3)Data3: Enron mail dataset; The pre processing used for Data1 was applied. This yielded a total set of 252487 mails and 106958 unique words.

(4)Data4: Netflix data; a set of rating records from Nov 1st, 1999 to Dec 31st, 2005. We first selected only those users who rated at least 20 movies and movies that were rated by at least 100 users. By focusing on the bias of ratings, we converted the ratings that were higher than the average rating of the user to 1 (purchase) and to 0 (no purchase) if not. Moreover, we rounded the day of grade (DD/MM/YYYY) to the month of grade (MM/YYYY). Each movie rating list of each user on the same month and the movies he rated are analogous to a document and the words in the document, respectively. This yielded a total set of 21033 movie rating lists (documents) from 90954 users and 7125 movies.

In our evaluation, the smoothing parameters α,β,γ and ϵ were set to {1/C(TDM,TAM)}, 1/Z, 0.1 and 1, respectively (all weak symmetric priors following previous works). We ran the experiments on PCs with Dual Core 2.66 GHz Xeon processors.

4.2 Qualitative Evaluation

4.2.1 Background topic

Table 2 demonstrates the background topic identified by TAM over all data sets. The background distributions of each set are quite similar to the appearance probability and so are intuitively interpretable, since words (movies) are commonly used (rated) across a broad range of documents (logs). Therefore, this result shows that TAM can distinguish the background topic. The ratio of background word and these words ranking remain basically constant regardless of the number of topics. Since Data2 covers more fields than Data1, its ratio of temporal words is larger than that of Data1 and its ratio of background words is fewer than that of Data1. The number of tokens in each document (mail) of Data3 is smaller than that of the others and so the ratio of background words is smallest among all data sets. The fact is that a huge number of users have commonly rated many of the same movies, while conference papers have more temporal and specific words.

4.2.2 Quality of trend class

Table 3 demonstrates the temporal words under the trend classes extracted by the TAM using the number of topics Z = 100 and trend class C = 100 from Data1 for each year. The temporal word distributions learned differs from the most likely words in the same year. As shown in the above row of this table, TAM assigns the likely words as background words, while the temporal words capture the trend of SIGIR conferences. "ranking" and 'indexing " are more popular words over time, "hypertext" and "web" became more popular after 1990. This table shows that TAM can identify the temporal words thanks to the trend class.

4.2.3 Trend class and Topic

Table 4 provides an example of the words under the trend class and topic learned by TAM from Data2. These selected trend classes have the most similar topic distribution, as measured by KL divergence eq (6), among all pairs of topic classes learned from Data2. As we can see, these trend class are commonly correlated with the topics of "statistics (ID 22)", and "information retrieval (ID 53)", where these titles are our interpretation from the most likely words associated with each topic. Conversely, these classes themselves are correlated with words describing the target of research. For example, trend class A (ID 2) is responsible for generating words that are used in network mining, multimedia web search, classification and summarization of Web data and so on, while trend class B (ID 7) is responsible for generating words that are used in knowledge extraction from documents and document structure. In fact, this interpretation is supported by the title list of papers associated with the corresponding trend class. Furthermore, the timestamps of these papers follow the continuous time distribution associated with each trend class. The time distribution of trend class A rises year after year, while that of trend class B is constant over the periods examined. These tables show that TAM can detect trends from the mixture of topics, temporal words and timestamps.

Table 2: Comparison of Background word on test sets: For each data set, we list 10 words(movies) by the appearance probability(above), and list 10 words(movies) by the learned back ground words probability under TAM(bottom) using (Z,C) = (100,100)(data1), (150,150)(data2), (170,170)(data3) and (120,120)(data4). The values in the columns are the corresponding probabilities.

Data1	Data2	Data3	Data4
retrieval 0.010	data 0.009	com 0.012	Miss Congeniality 0.002
information 0.010	query 0.006	enron 0.008	Independence Day 0.002
query 0.010	set 0.006	subject 0.004	The Patriot 0.002
documents 0.009	information 0.005	new 0.004	The Day After Tomorrow 0.002
document 0.008	web 0.005	http 0.004	Pirates of the Caribbean 0.002 The Curse of the Black Pearl
terms 0.006	number 0.005	power 0.004	Pretty Woman 0.002
set 0.005	using 0.005	sent 0.004	Forrest Gump 0.002
model 0.005	model 0.004	www 0.003	The Green Mile 0.002
search 0.005	figure 0.004	energy 0.003	Con Air 0.002
used 0.005	algorithm 0.004	said 0.003	Twister 0.002
information 0.001	acm 0.001	subject 0.004	The Day After Tomorrow 0.002
retrieval 0.001	references 0.001	com 0.003	Miss Congeniality 0.002
acm 0.001	abstract 0.001	know 0.003	The Sixth Sense 0.002
used 0.001	information 0.001	enron 0.003	Armageddon 0.002
using 0.001	introduction 0.001	sent 0.003	The Rock 0.002
use 0.001	use 0.001	items 0.003	Forrest Gump 0.002
search 0.001	work 0.001	thanks 0.003	I, Robot 0.002
references 0.001	copyright 0.001	folderssent 0.003	The Italian Job 0.002
results 0.001	using 0.001	original 0.003	The Bourne Identity 0.002
number 0.001	terms 0.001	pmto 0.002	Lost in Translation 0.002

Table 3: Word distribution of temporal topic from Data1:We list 10 words by their appearance probability from each year(above), and list words by the learned probability from the topic class under TAM(bottom) over 25 years. A topic class is assigned to each year's data by using the continuous time distribution associated with each trend class. The values in the columns are the corresponding probabilities.

1980	1985	1990	1995	2000	2005
information 0.017	document 0.012	query 0.012	documents 0.011	information 0.012	retrieval 0.009
retrieval 0.010	documents 0.012	document 0.011	retrieval 0.010	documents 0.011	information 0.009
document 0.007	query 0.011	retrieval 0.011	document 0.009	document 0.009	query 0.008
search 0.007	retrieval 0.011	information 0.011	information 0.009	retrieval 0.009	documents 0.007
time 0.006	information 0.010	documents 0.009	query 0.008	query 0.006	model 0.007
documents 0.006	terms 0.007	terms 0.009	set 0.007	used 0.006	document 0.007
query 0.006	number 0.006	search 0.007	terms 0.007	search 0.006	data 0.006
terms 0.006	queries 0.006	term 0.006	used 0.006	using 0.005	results 0.006
set 0.006	set 0.006	number 0.006	text 0.005	results 0.005	search 0.006
data 0.006	search 0.006	set 0.006	number 0.005	text 0.005	set 0.006
systems 0.006	data 0.006	used 0.006	queries 0.005	number 0.005	using 0.005
similarity 0.226	model 0.282	thesaurus 0.167	database 0.185	relevance 0.152	web 0.225
ranking 0.193	text 0.263	collection 0.151	indexing 0.173	relevant 0.112	relevance 0.215
initial 0.082	indexing 0.182	function 0.127	effectiveness 0.162	topics 0.107	topic 0.176
mapping 0.057	database 0.176	order 0.112	model 0.154	database 0.104	feedback 0.153
feature 0.056	language 0.128	boolean 0.092	index 0.131	filtering 0.098	trec 0.146
representative 0.048	structure 0.113	vector 0.088	algorithm 0.096	process 0.088	ranking 0.142
supported 0.046	matching 0.104	expert 0.082	network 0.088	ranking 0.075	music 0.121
decision 0.039	descriptions 0.096	systems 0.065	ranking 0.075	recall 0.068	results 0.105
assignment 0.035	n-gram 0.083	vertices 0.059	structure 0.069	selection 0.058	ranking 0.093
data 0.033	ranking 0.077	hypertext 0.052	clustering 0.067	web 0.042	set 0.052
larger 0.032	systems 0.069	online 0.043	tree 0.057	databases 0.037	filtering 0.032

Table 4: Word distribution of trend class and topics from Data2: We show the continuous time distribution (left column), and list 10 words by the learned probability from the trend class under TAM (middle column), and list 10 words by the learned probability from the topic under TAM (right 4 columns)using the number of topics $Z = 100$ and trend class $C = 100$. Each topic is decided by highest value of the topic distribution associated with each trend class.

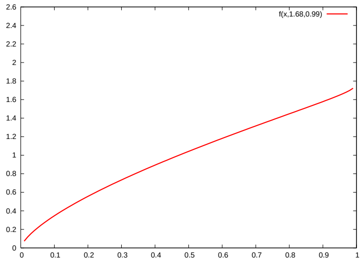
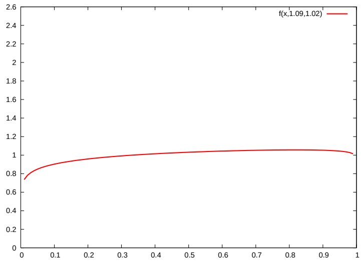
Time	Trend	Topic			
	ID 2	ID 22	ID 29	ID 53	ID 82
	web	statistical	machine	recall	algorithm
	network	parameter	personalized	precision	clustering
	image	evaluation	model	algorithms	node
	hybrid	smoothing	learning	sentences	nodes
	community	estimating	training	similarity	cluster
	node	probabilities	preference	equivalent	kernel
	annotation	sampling	topic	accurate	hierarchical
	digital	heuristic	functions	frequency	classification
	color	bayesian	communities	relevant	singular
	objects	variant	filtering	index	subspace
	feature	threshold	collaborative	query	tree
Time	ID 7	ID 8	ID 22	ID 53	ID 71
	news	discovery	statistical	recall	linear
	database	mining	parameter	precision	svm
	initial	patterns	evaluation	algorithms	optimization
	xml	algorithm	smoothing	sentences	matrix
	sentence	tuples	estimating	similarity	unlabeled
	learn	sequence	probabilities	equivalent	kernel
	knowledge	computational	sampling	accurate	theorem
	decision	detection	heuristic	frequency	classification
	dictionary	sequence	bayesian	relevant	singular
	bilingual	implemented	variant	index	regression
	structure-based	finding	threshold	query	semi-supervised

Table 5: Average KL divergence between word distributions learned on data2: All models are learned with the number of topics Z set at 150 and the number of trend classes C set at 150. Results that differ significantly t-test $p < 0.01$, $p < 0.05$ from TOT are marked with '', '*' respectively.**

Distance	TOT	sTOT	TAM
$D_{KL}(\phi_z \mid \phi_{z'})$	16.3	16.5	16.8^{**}
$D_{KL}(\phi_z \mid \phi_b)$	-	11.0	13.1
$D_{KL}(\phi_z \mid \phi_c)$	-	-	11.8
$D_{KL}(\phi_c \mid \phi_{c'})$	-	-	4.8
$D_{KL}(\phi_c \mid \phi_b)$	-	-	6.7

4.3 Quantitative Evaluation

4.3.1 *Effect of switch variable*

We evaluate average word distribution separations between all pairs of different classes and discuss the effect of the switch variable in tracking trends. We measure this distance between topics by the average KL-Divergence. Table 5 shows the results of the distance comparison. From this table, we observe the following: (1)Distinct topics: Comparisons on ϕ_z show that the topics yielded by TAM have higher score. This shows that TAM can identify more distinct topics than the other models. (2)Difference of words: The TAM results show that the score for different classes is as large as that within same class. This implies that background words and temporal words differ distinctly from topic words.

4.3.2 *Effect of trend class*

To measure the ability of the proposed models to act as generative models, we computed test-set perplexity under the estimated parameters and compared the resulting values. Perplexity, which is widely used in the language modeling community to assess the predictive power of a model, is algebraically equivalent to the inverse of the geometric mean per-word likelihood (lower numbers are better). The perplexity was computed for all algorithm using 100 samples from 100 different chains using

$$PPX = \exp\left(-\frac{1}{W}\sum_{d \in D_{\text{test}}}^{|D_{\text{test}}|}\sum_{v \in d}^{|D|}\frac{1}{G}\log\left(\sum_{z}^{Z}\theta_z^g \phi_{zv}^g\right)\right), \quad (7)$$

where W is the number of test words, G is the number of samples (from G different chains), θ_z^g is the probability that z will be assigned by a model to document d in g and ϕ_{zv}^g is the probability assigned by the model to word v conditioned on z in g. A lower score implies that word w_d is less surprising to the model. Note that TDM and TAM sample topics from the trend class assigned in each document. Since TDM incorporates the switch variable in it, we modify eq (7) as the follows:

$$PPX = \exp(-\frac{1}{W}\sum_{d\in D_{\mathrm{test}}}^{|D_{\mathrm{test}}|}\sum_{v\in d}^{|D|}\frac{1}{G}\log(\mu_0^g\phi_{bv}^g+\mu_2^g\phi_{cv}^g+\sum_{z}^{Z}\mu_1^g\theta_{cz}^g\phi_{zv}^g)), \quad (8)$$

where μ_0^g (μ_1^g, μ_2^g) is the probability of switch variable in each token μ_0 (μ_1, μ_2) in g, θ_{cz}^g is the probability that z will be assigned by a trend class c in document d in g, and ϕ_{bv}^g is the probability assigned by the model to word v conditioned on background topic b in g, and ϕ_{cv}^g is the probability assigned by the model to word v conditioned on c in g.

We computed the perplexity as follows. First, we randomly split 10% of each document to create a test part; the remainder was used as the learning part. For every document, the test part was held out to compute perplexity. Second, the learning part was used for estimating the parameters by Gibbs sampling. Finally, a single set of topic counts was saved when a sample was taken; the log probability of test words that had never been seen before was computed in the same way as the perplexity computation of previous works. We fixed the number of topics (Z,C) = (100,100)(data1), (150,150)(data2), (170,170)(data3) and (120,120)(data4) for simplicity and fair comparison. In this calculation, we removed low frequency words (movies), those that appeared in fewer than 3% of all documents (lists), as stop words. We selected the slice size of time as year (data1 and data2), and month (data3 and data4).

Table 6 shows the results of the perplexity comparison. These results are averaged over five-fold cross validation. From this table, we observe the following: (1)Dominant words: The results over all data sets, that TAM offers lower perplexity than TOT, TAM with limited r, support our idea that background and temporal words exist in each document and lead to a reduction in perplexity. Although the ratio of these words to terminology words varies with the data set, detecting these differences is required for modeling the topic evolution precisely. (2)Noise reduction: The trend class allows TAM to group documents under the various topic distributions rather than permitting various topic distributions on each document. Consequently, documents placed in the same trend class have the same topic distribution. This implies that clustered documents contain less noise than otherwise, and so reduce overall perplexity.

To discuss the effect of the trend class, we reran the perplexity comparison with different settings. Since it is difficult to compare word distributions and t simultaneously by the same criterion, we used PPX and L1 error to compare TAM and TOT by varying both $|C|$ and $|Z|$. Table 7 shows that many words tend to be generated from trend class, since the ratio of $r = 2$ increases in proportion to the number of trend classes $|C|$ and of topics $|Z|$.

Figure 2 shows that TAM attains lower perplexity than TOT under the condition that the total number $|C| + |Z|$ of TAM is smaller than the number $|Z|$ of TOT, in many cases, even if TAM(C) is shifted $|C|$ to the right. For example, TAM($|C|$=50 and $|Z|$=100) attains a lower perplexity than TOT($|Z|$=200). From this figure, we observe the following: (1)Dominance of trend class: The decrease of perplexity slows in inverse proportion to the increase of number of topics, since an increase in the number of topics does not significantly affect topic assignment; thus assignment should be relatively invariant to an increase in topics [17]. On the contrary, the decrease of perplexity rises in inverse proportion to the increase of number of trend class. Since TAM groups documents of similar topics by using the trend class, words tend to be generated from the trend class directly rather than indirectly from topics. Consequently, this result shows that the change in perplexity is more sensitive to the number of trend classes than that of topics. (2)The number of trend/topic classes: In each document, about 80% of all tokens are occupied by a few kind of topic class and many of these topics are reused in other documents. Consequently, we need trend classes more than topics, for tracking trends. Experiments shows that the deficient in TOT can be conquered by the incorporation of both c and r into TOT and changing the position of t in a document.

Additionally, we compared the change in perplexity over the number of Gibbs sampling; the results are shown in Figure 3. From this figure, we observe the following: (1)Temporality: Given the same number of iterations, increasing

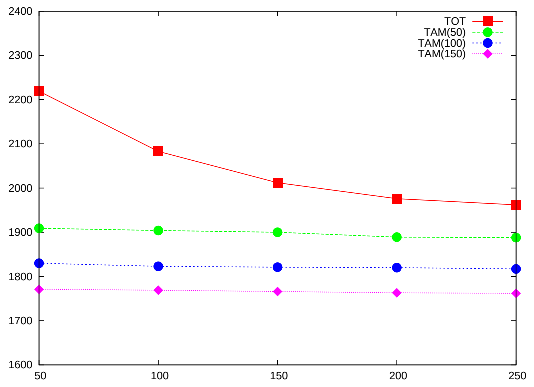


Figure 2: Perplexity comparison of TOT and TAM run with different settings of topic Z and trend class C on Data2: In this figure, the x-axis denotes the Z value and the y-axis denotes the perplexity value. The number in parenthesis denotes the C value.

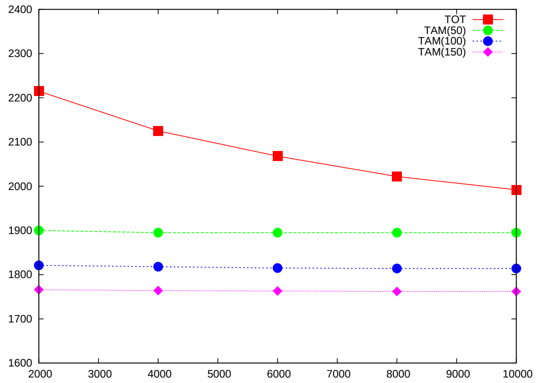


Figure 3: Perplexity comparison of TOT and TAM run on different settings of the iterations in Gibbs sampling on Data2: In this figure, the x-axis denotes the number of iterations in Gibbs sampling and the y-axis denotes the perplexity value. The number in parenthesis denotes the C value. Z was set to 150 for all models

Table 6: Perplexity comparison of TOT, sTOT and TAM on test sets: The number in the second row for TAM is the number of trend classes. The sets in the second row for TAM denote the sets of available switch r in TAM(r={0,1,2}), and the corresponding set of numbers is the ratio of r(r=0:r=1 and r=0:r=1:r=2). Results that differ significantly t-test $p < 0.01$, $p < 0.05$ from TOT are marked with '', '*' respectively.**

Data	TOT	sTOT	TAM		
			$\{r = 1\}$	$\{r = 0 : r = 1\}$	$\{r = 0 : r = 1 : r = 2\}$
Data1	1528	1438^{**}	1483^{*}	1419^{**} $\{32.7 : 67.3\}$	1356^{**} $\{22.6 : 31.8 : 45.6\}$
Data2	2219	1982^{**}	2167^{*}	2089^{**} $\{31.6 : 68.4\}$	1766^{**} $\{27.8 : 32.2 : 40.0\}$
Data3	2225	2176^{*}	2155^{**}	2132^{**} $\{0.5 : 99.5\}$	2116^{**} $\{0.4 : 35.1 : 64.5\}$
Data4	1203	1032^{**}	996^{**}	988^{**} $\{49.2 : 50.8\}$	966^{**} $\{27.2 : 40.5 : 32.3\}$

Table 7: The ratio of switch variable and PPX comparison of TAM on Data2

Trend	|C|=50			|C|=100			|C|=150		
Topic	|Z|=50	|Z|=100	|Z|=150	|Z|=50	|Z|=100	|Z|=150	|Z|=50	|Z|=100	|Z|=150
r=0	27.2	33.7	32.3	32.6	33.1	36.2	36.6	28.4	27.8
r=1	40.6	34.8	30.2	37.7	37.8	28.2	32.0	34.2	32.2
r=2	32.2	31.5	37.5	29.7	29.1	35.6	31.4	37.4	40.0
PPX	1909	1902	1900	1830	1823	1821	1771	1769	1766

Table 8: L1 comparison of TOT and TAM: All models are learned with (Z,C) = (100,100)(data1), (150,150)(data2), (170,170)(data3) and (120,120)(data4). Results that differ significantly by parametric nonpaired t-test $p < 0.01$, $p < 0.05$ from TOT are marked with '', '*' respectively.**

Data	Data1	Data2	Data3	Data4
TOT	2.44	2.25	2.11	1.57
TAM	2.17^{**}	2.03^{**}	1.88^{**}	1.16^{**}

the number of trend classes is more effective in reducing the value of perplexity than increasing that of topics, like the characteristic shown in the previous figure. Moreover, TAM tends to achieve low perplexity faster than TOT and this fall is also proportional to the number of trend classes. This implies that not only words but also topics exhibit temporality. (2)Abridgement: TAM achieves low perplexity with fewer iterations than TOT. Since the trend class can abridge topics in each document using θ_{TAM} associated with trend class, TAM handles both temporal and terminology words in each document in a practical manner by using the trend class. Consequently, the increase in the number of trend classes more significantly impacts perplexity than that of topics.

4.3.3 *Time prediction*

One interesting common feature of both TOT and TAM is their ability to predict the timestamp given the words in a document. This functionality also provides another opportunity to quantitatively compare TAM against TOT. On all data sets, we measure the ability to predict the year given data set 1 and data set 2, month given data set 3 and data set 4, as measured by accuracy, L1 error. As shown in Table 8, TAM achieves double the accuracy of TOT, and provides an L1 relative error reduction of 20% on average. Although each document has only one time stamp, TOT would generate different time stamps within the same document. Consequently, 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, TAM attains lower perplexity and more distinct topics than TOT for identifying low dimensional components.

5. DISCUSSION

One of the most significant differences between TAM and TOT is the ability to provide a fully generative model. Since TOT creates each topic with both a time distribution and a word distribution, it separates the sets of same words that have different timestamps by assigning them to different topic classes. Thus, a tunable hyper-parameter is required to prevent timestamp generation from being overwhelmed by the multiplicity of words generated under the bag of words assumption. TAM handles temporal words, words associated with constant topics and background class words through the value of **r**; it can learn distinct word distributions with fewer latent variables than TOT, where each topic generates words constantly over time. Accordingly, TAM predicts a topic distribution as conditioned by the timestamp given by this trend class, while TOT predicts this by using a Bayes rule on topics.

Note that whether each word is temporal or not depends on the timespan of the given corpus. For example, "Internet" is a temporal word in the news archives of the last 100 years, while "Internet" is a non-temporal word in the acm data set of the past 10 years.

The advantage of TAM over TOT is in reducing the cost of calculation(CPU). TAM learns the word distribution from all documents with similar topics and topic distributions by merging similar distributions into one distribution, while TOT estimates the topic distribution at the document level without using the intersimilarity of documents. Accordingly, TOT needs more latent topic variables Z for learning the distinct word distribution than TAM as shown in Table 5. As shown in Figure 2, TAM gains low perplexity with just half the number of topics needed by TOT. Since $|C_{TAM}| \approx |Z_{TOT}|$ is enough for tracking trends as stated in the effect of trend class, TAM achieves lower perplexity with few iterations and topics than the others, and offers significantly lower perplexity under the same condition on the number of topics with TOT, see Figure 2 and Figure 3.

6. CONCLUSION

In this paper, we proposed a topic model that explicitly models time jointly with temporal word co occurrence and topic co-occurrence. A novel feature of our model is the inclusion of trend class into a topic model; this class is responsible for generating both observed timestamps, and the sets of topics as well as words, where each topic is responsible for generating words. Moreover the switch variable, also included in the model, can distinguish these different types of words in each token of each document. Experiments on various data sets showed that the proposed model can capture interpretable low dimensionality sets of topic class, and is useful for analyzing the evolution of trends. In future work, we will extend TAM by incorporating other metadata, and then apply this model to collaborative filtering focused on the dynamics of user preference [6].

7. REFERENCES

[1] D. Blei and J. Lafferty. Dynamic topic models. 23:113–120, 2006.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[3] D. Blei, T.Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. *NIPS*, 16, 2004.

[4] A. Gohr, A. Hinneburg, R. Schult, and M. Spiliopoulou. Topic evolution in a stream of documents. In *SDM*, pages 859–872, 2009.

[5] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD*, pages 168–177, 2004.

[6] N. Kawamae. Latent interest-topic model:finding the conditional relationships between author-document. In *CIKM*, pages 649–658, 2010.

[7] N. Kawamae. Serendipitous recommendations via innovators. In *SIGIR*, pages 218–225, 2010.

[8] N. Kawamae. Predicting future reviews: Sentiment analysis models for collaborative filtering. In *WSDM*, page To appear, 2011.

[9] N. Kawamae and R. Higashinaka. Trend detection model. In *WWW*, pages 1129–1130, 2010.

[10] J. Kleinberg. Bursty and hierarchical structure in streams. In *KDD*, pages 91–101, 2002.

[11] J. Konstan, B. Miller, J. H. D. Maltz, L. Gordon, and J. Riedl. Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

[12] Y. Koren. Collaborative filtering with temporal dynamics. In *KDD*, pages 447–456, 2009.

[13] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML*, pages 577–584, 2006.

[14] Q. Mei and C. Zhai. A mixture model for contextual text mining. In *KDD*, pages 649–655, 2006.

[15] R. Nallapati, W. Cohen, S. Ditmore, J. Lafferty, and K. Ung. Multi-scale topic tomography. In *KDD*, pages 520–529, 2007.

[16] A. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *EMNLP*, 2005.

[17] H. Wallach, D. Mimno, and A. McCallum. Rethinking lda: Why priors matter. In *NIPS*, pages 859–872, 2009.

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

[19] L. Zhuang, F. Jing, and X. Zhu. Movie review mining and summarization. In *CIKM*, pages 43–50, 2006.

APPENDIX

As shown in eq (1), multinomials can be adapted by the conjugate prior and then integrated out analytically as follows:

$$
\begin{aligned}
&p(\mathbf{w},\mathbf{z},\mathbf{r},\mathbf{t},\mathbf{c}|\alpha,\beta,\gamma,\lambda,\epsilon) = \int p(\mathbf{w},\phi|\mathbf{z},\gamma)d\Phi \int p(\mathbf{z},\theta|\mathbf{r},\mathbf{c},\beta)d\Theta \\
&\times \int p(\mathbf{r},\mu|\epsilon)d\mu p(\mathbf{t}|\mathbf{c},\lambda) \int p(\mathbf{c},\psi|\alpha)d\Psi \\
&= \int \prod_d^D P(c_d|\psi_{a_d})d\Psi \prod_a^A p(\psi_a|\gamma) \\
&\times \int \prod_d^D \prod_i^{N_d} P(z_{di}|c_d,\theta_{c_d}) \prod_j^J p(\theta_j|\delta)d\Phi \\
&\times \int \prod_d^D \prod_i^{N_d} P(w_{di}|z_{di},\phi_{z_{di}}) \prod_t^T p(\phi_t|\beta)d\Phi \\
&= \int \prod_a^A \frac{\Gamma(\sum_j^J \gamma_j)}{\prod_j^J \Gamma(\gamma_j)} \prod_j^J \psi_{aj}^{\gamma_j-1} \prod_a^A \prod_j^J \psi_{aj}^{n_{aj}} d\Psi \\
&\times \int \prod_j^J \frac{\Gamma(\sum_t^T \delta_t)}{\prod_t^T \Gamma(\delta_t)} \prod_t^T \theta_{jt}^{\delta_t-1} \prod_j^J \prod_t^T \theta_{jt}^{n_{jt}} d\Theta \\
&\times \int \prod_t^T \frac{\Gamma(\sum_v^V \beta_v)}{\prod_v^V \Gamma(\beta_v)} \prod_v^V \phi_{tv}^{\beta_v-1} \prod_t^T \prod_v^V \phi_{tv}^{n_{tv}} d\Phi \\
&= [\frac{\Gamma(\sum_c^C \alpha_c)}{\prod_c^C \Gamma(\alpha_c)}][\frac{\Gamma(\sum_r^R \epsilon_r)}{\prod_r^R \Gamma(\epsilon_r)}]^D [\frac{\Gamma(\sum_z^Z \beta_z)}{\prod_z^Z \Gamma(\beta_z)}]^C [\frac{\Gamma(\sum_v^V \gamma_v)}{\prod_v^V \Gamma(\gamma_v)}]^{Z+C+1} \\
&\times \frac{\prod_c^C \Gamma(n_c+\alpha_c)}{\Gamma(\sum_c^C (n_c+\alpha_c))} \prod_c^C [\frac{\prod_z^Z \Gamma(n_{cz}+\beta_z)}{\Gamma(\sum_z^Z (n_{cz}+\beta_z))}] \prod_d^D p(t_d|\lambda_{c_d}) \\
&\times \prod_d^D [\frac{\prod_r^R \Gamma(n_{dr}+\epsilon_r)}{\Gamma(\sum_r^R (n_{dr}+\epsilon_r))}] \prod_z^{Z+C+1} \frac{\prod_v^V \Gamma(n_{zv}+\gamma_v)}{\Gamma(\sum_v^V (n_{zv}+\gamma_v))},
\end{aligned}
\tag{9}
$$

where n_c represents the number of documents assigned to trend class, c, $n_{cz}(n_{zv})$ represents the number of topics, z(word v), assigned to trend class c(topic z), and n_{dr} represents the number of switch variables, r, assigned to document d.