Infinite Topic Modelling for Trend Tracking
Hierarchical Dirichlet Process Approaches with Wikipedia Semantic based Method

Yishu Miao[1], Chunping Li[1], Hui Wang[2] and Lu Zhang[1]
[1]*School of Software, Tsinghua University, Beijing, China*
[2]*School of Computing and Mathematics, University of Ulster, Jordanstown, U.K.*

Keywords: Hierarchical Dirichlet Process, Topic Modelling, Wikipedia, Temporal Analysis, News.

Abstract: The current affairs people concern closely vary in different periods and the evolution of trends corresponds to the reports of medias. This paper considers tracking trends by incorporating non-parametric Bayesian approaches with temporal information and presents two topic modelling methods. One utilizes an infinite temporal topic model which obtains the topic distribution over time by placing a time prior when discovering topics dynamically. In order to better organize the event trend, we present another progressive superposed topic model which simulates the whole evolutionary processes of topics, including new topics' generation, stable topics' evolution and old topics' vanishment, via a series of superposed topics distribution generated by hierarchical Dirichlet process. Both of the two approaches aim at solving the real-world task while avoiding Markov assumption and breaking the number limitation of topics. Meanwhile, we employ Wikipedia based semantic background knowledge to improve the discovered topics and their readability. The experiments are carried out on the corpus of BBC news about American Forum. The results demonstrate better organized topics, evolutionary processes of topics over time and model effectiveness.

1 INTRODUCTION

At the outset of this work lies the observation that in the analysis of time stamped documents, such as news corpus, people are concerned about what events have taken place and their entire evolutionary processes. As we can see from Figure 1, each colour represents a trend that accords with the timestamp on the timeline above. Correspondingly, there exist several articles related to each trend, such as *"Media condemn N Korea nuclear test"* in May.26th, *"North Korea increases its leverage"* in Jun.8th, *"Obama 'prepared' for N Korea test"* in Jun.21st and *"Clinton's high drama Korean mission"* in Aug.6th. The successive articles indicate a period of attention paid by Americans worrying about Korea nuclear issue. Despite each article differs in the title, they are concerning the same topic. Hence, we attempt to discover the latent topic via topic modelling on the content of these articles .

Topic modelling (Landauer, 1997) (Hofmann, 1999) has been a prevailing method on text analysis and feature reduction. The topics are called "reduced description"(Blei et al., 2003) associated with the documents. Thus, we extract the trend via bunches of words with probability represented by topic. As time elapses, some topics which drew people's attention will fade away from public view, while some new topics may raise a great awareness conversely accompanying with the occurrence of big events. Moreover, there also exist several topics which attract persistent attention and become significant parts of people's daily life. All of these topics are crucial to trend tracking.

The purpose of trend tracking focuses on temporal information and related entities. (Blei and Lafferty,), (Wang et al., 2008), (XueruiWang and McCallum, 2006) and (AlSumait et al., 2008), etc. are those topic models extracting the temporal information while modelling topic distribution. However, considering the compatibility of Markov assumption and Dirichlet distribution, it is not amenable to use sequential modelling method in traditional topic models with Dirichlet prior. After deliberating the generation of topics when applying non-parameter Bayesian approach, we find that there exists significant temporal information not only in the words of documents, but also in the generative process of topics. Hence, we associate the models with the time information and

Miao Y., Li C., Wang H. and Zhang L..
Infinite Topic Modelling for Trend Tracking - Hierarchical Dirichlet Process Approaches with Wikipedia Semantic based Method.
DOI: 10.5220/0004133300350044
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2012), pages 35-44
ISBN: 978-989-8565-29-7

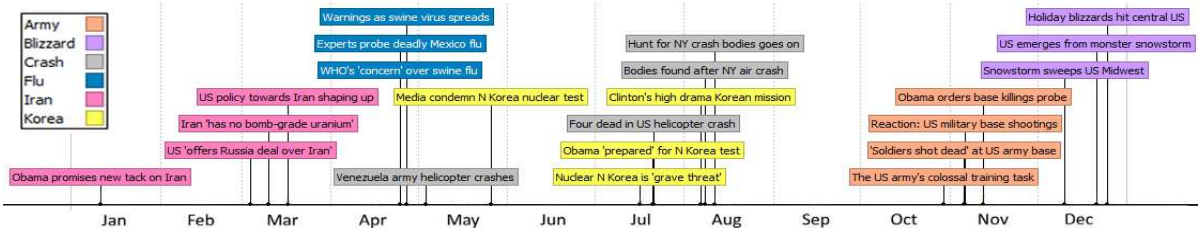


Figure 1: Articles on timeline of year 2009.

no Markov assumption in two approaches: to model a distribution over time, or to make use of the dynamic feature of Hierarchical Dirichlet process(Teh et al., 2006) alternatively. When applied to tackle the real-world task such as trend tracking, the models without Markov assumption are much easier not only when building a succinct graphic model but also to depress the complexity of inference algorithm.

Firstly, we employ a time variable in the first approach to model the evolutionary process of topic based upon the publishing time of documents while discovering topics. As an infinite topic model, it is more likely to generate a new topic at the timestamp where the texts congregate together. From the Chinese restaurant process perspective(Sudderth, 2006), a man prefers to sit on a new table if he knows that at 12:00, there will be many clients coming in, for he doesn't want to have dinner with too many strangers.

Then, we present a superposition topic model which generates new topics progressively. In this model, at first, we divide the corpus into several parts in the time order. After we achieve the topic distribution over terms on the first sub-corpus, it has already generated corresponding topics without setting the number of the topics beforehand. Based on the previous distributions, this model assigns the terms to join the preceding topics or hold together in groups as new topics. Besides, those topics, seldom discussed in the following corpus, will vanish after several iterations as vanished topics. While a majority of topics will be stable topics in their whole evolutionary processes due to persistent attention of public such as *Obama* and *Military* in American news.

After introducing the two infinite topic modelling approaches, we consider incorporating a semantic method to improve the experiment results. Even the conventional media, i.e. newspaper, incline to illustrate an event via hackneyed way which leads to a small moiety of trivial words and exaggerate rhetoric. In order to tackle the impediment, we attempt to employ Wikipedia semantic background knowledge to improve the readability of discovered topics. By mapping the terms in the articles to Wikipedia concepts, we will achieve a majority of entities from the corpus to improve the granularity of extracted topic distribution. Hence, we build an entity model to analyze the event trend based on the means mentioned above. In this model, we will not discard the words after mapping them to entities, but sample every one during Gibbs Sampling process and it will contribute to the generation of topic distribution over entities.

The experiments and evaluation are mainly on the BBC news of American forum which contains normative articles and precise publishing time. In Section 2, we review some related developments of infinite topic models and temporal information analysis. Then, we introduce two non-parametric Bayesian approaches and the entity model based on Wikipedia semantic method in Section 3.In Section 4, we present the experiment result analysis according to the comparison of different models. Finally, we have the concluding remarks and future work in Section 5.

2 RELATED WORK

In this section, we briefly introduce related work including non-parametric topic modelling methods, temporal information analysis and semantic knowledge based method.

Basically, hierarchical Dirichlet process(HDP) (Teh et al., 2006) as an extended Dirichlet process(DP)(Ferguson, 1973) is a typical implementation of non-parametric Bayesian approach. Such as infinite LDA(Heinrich, 2011), it achieves a relative satisfactory result with a low level of complexity. Besides, dHDP(Ren et al., 2008) assumed that each paragraph of one document is associated with a particular topic, and the probability of a given topic being manifested in a specific document evolves dynamically between contiguous time-stamped documents. But it only presents the infinity of time-stamp number, while the topic number is still a limitation in the topic evolutionary scenario. Evo-HDP(Zhang et al., 2010) and infinite Dynamic topic models(Ahmed and Xing, 2010) can automatically determine the topic number, but both are based on Markov assumption. (Balasubramanyan et al., 2009) is similar to our first approach, which combines HDP and TOT model simply, but

makes no use of time information when generating new topic. Besides, there also exist an approach about trend tracking via a combination of Dynamic topic models and time series methods without HDP(Hong et al., 2011).

Semantic based methods are usually used for the purpose of analysing short textual data or eliminating multilingual and ambiguous problems. There exist several approaches incorporating topic modelling. For example, (Kataria et al., 2011) uses Wikipedia annotations to label the entities and learn word-entity associations. (Ni et al., 2009) takes advantage of multilingual corpus of Wikipedia to extract the universal topics and cluster the terms of different languages. Besides, the plentiful context information of Wiki concepts is profitable to automatic topics labelling. (Lau et al., 2011) presents a typical approach to tackle the topic comprehension problem caused by unsupervised topic modelling. However, we employ Wikipedia as background knowledge in this scenario aiming at improving the quality of topic clustering which is different from these models mentioned above. Based on CorrLDA(Newman et al., 2006), we discover the topics distribution over entities extracted by Wikipedia without discarding words when sampled in inference step and improves the readability of topics entirely.

Without Markov assumption, our approach is considered more succinct and easy to be employed on the real-world task when compared to other existing methods. In this paper, we also make a comparison of these two approaches and have discovered the different features between them, which will be presented in the experiment part.

3 INFINITE TOPIC MODELLING FOR TREND TRACKING

At the beginning of this section, we briefly introduce Dirichlet process mixtures and HDP. Then the two infinite topic modelling approaches, including basic instruction, graphic representation and inference process, are discussed in the following part. Afterwards, we present the Wikipedia based semantic approach and its implementation method.

3.1 Hierarchical Dirichlet Process and Non-parametric Graphic Model

DP have been used as non-parameter Bayesian approach to estimate the number of components which define a distribution over distributions. We use $G \sim DP(\alpha_0, G_0)$ to represent a DP, where G_0 is base measure, and α_0 as the concentration parameter is a positive real number. Dirichlet mixture model, employ DP as a non-parameter prior on the latent parameter distribution, is one of the most significant applications of DP. In Dirichlet process mixtures, we suppose:

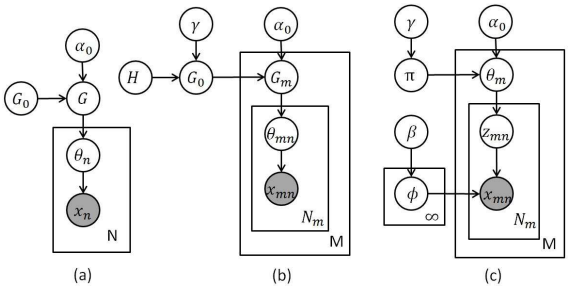


Figure 2: (a) Dirichlet process, (b) Hierarchical Dirichlet process, (c) Stick-breaking representation of HDP.

$$\theta_m|G \sim G$$
$$x_m|\theta_m \sim F(\theta_m) \quad (1)$$

θ_m denotes the parameter of *mth* component, while $F(\theta_m)$ denotes the distribution when given θ_m and Figure 2(a) shows the graphic representation. From the perspective of Stick-breaking construction, we place a Dirichlet process prior on the latent parameter distribution $G \sim DP(\alpha, H)$. Hence the other representation of a Dirichlet process mixture is presented as follows:

$$\boldsymbol{\pi}|\alpha_0 \sim GEM(\alpha_0) \qquad z_n|\boldsymbol{\pi} \sim \boldsymbol{\pi}$$
$$\theta_k|G_0 \sim G_0 \qquad x_n|\theta_{z_n} \sim F(\theta_{z_n}) \quad (2)$$

The random probability distribution G on θ satisfies $G(\theta) = \Sigma_{k=1}^{\infty}\pi_k\delta(\theta, \theta_k)$. In this formula, π, as a random probability measure on positive integers, satisfies $\Sigma_{k=1}^{\infty}\pi_k = 1$. GEM stands for Griffiths, Engen and McCloskey and the construction of $\pi|\alpha_0 \sim GEM(\alpha_0)$ can be presented in the Stick-beaking construction as follows:

$$\pi_k^{'}|\alpha_0 \sim Beta(1, \alpha_0)$$
$$\phi_k|G_0 \sim G_0 \quad (3)$$

Then we define a random measure G as:

$$\pi_k = \pi_k^{'}\prod_{l-1}^{k-1}(1-\pi_l^{'})$$
$$G = \Sigma_{k=1}^{\infty}\pi_k\delta_{\phi_k} \quad (4)$$

Hence, it is an alternative way to express Dirichlet mixture model.

Nevertheless, no sharing can occur between groups of data if a single Dirichlet process is applied. In order to link these mixture models, the base distribution can be itself distributed as a Dirichlet process,

and then it allows groups share statistical strength. This non-parameter Bayesian approach is hierarchical Dirichlet process(Teh et al., 2006). A HDP defines a set of random probability measures G_j for group j and each of them is drawn from a $DP(\alpha_0, G_0)$. Moreover, the global measure G_0 is also drawn from a $DP(\gamma, H)$. The definition can be presented as follows, and Figure 2(b) shows the graphical model:

$$G_0|\gamma, H \sim DP(\gamma, H) \quad G_m|\alpha_0, G_0 \sim DP(\alpha_0, G_0)$$
$$\theta_{mn}|G_m \sim G_m \quad x_{mn}|\theta_{mn} \sim F(\theta_{mn}) \qquad (5)$$

The base measure H is drawn from a DP, hence every child shares the measure and is conditionally independent with each other. Correspondingly, we give the Stick-breaking construction perspective and the graphic model in Figure 2(c).

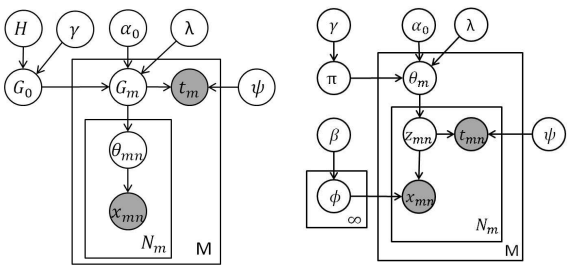
$$\boldsymbol{\pi}|\gamma \sim GEM(\gamma)$$
$$\theta_m|\alpha_0, \boldsymbol{\pi} \sim DP(\alpha_0, \boldsymbol{\pi}) \quad z_{mn}|\theta_m \sim \theta_m$$
$$\phi_k|H \sim H \quad x_{mn}|\phi_{z_{mn}} \sim F(\phi_{z_{mn}}) \qquad (6)$$

As illustrated above, all the measures are drawn from the base DP, which means that the discrete base distribution is shared by every document in the corpus. The global cluster weighs $\boldsymbol{\pi} \sim GEM(\gamma)$ follow a Stick-breaking process and denote the Dirichlet prior of infinite topic distribution over terms.

3.2 Infinite Temporal Topic Model

In this section, we present infinite temporal topic model(ITTM) to incorporate time information during the topic discovering based on the infinite model(Heinrich, 2011).

Ordinarily, HDP is used as prior in the mixture models to create infinite topic model and limited in a specific time period. Virtually, the universal time assignment on documents could be taken advantage of, and used as a time prior on probability measures when be drawn from global measure via DP. To put it another way, it obtains a higher probability to generate a new topic when the words congregate together at this timestamp in Dirichlet Process. Similar to Topics over Time model, ITTM avoids discretization by associating continuous time distributions with topics. The time range is normalized between 0 and 1 so that the Beta distribution can easily form the peaks of every time distribution of topics based on the time variable.



(a) Bayesian network (b) Stick-Breaking representation

Figure 3: Infinite Temporal Topic Model.

As illustrated in Figure 3, the ITTM can be represented in two perspectives. λ is the parameter of Beta distribution on time prior. t_{mn} represents the timestamp associated with word n in document m and ψ denotes its parameter of Beta distribution. The generative process is described as follows:

1. Draw an infinite dimension multinomial $\boldsymbol{\pi}$, $\boldsymbol{\pi}|\gamma \sim GEM(\gamma)$
2. For each toic z, draw a multinomial ϕ_z, $\phi_{z_{mn}}|\beta \sim Dirichlet(\beta)$
3. For each document m,
   - *3.1 Draw a time prior* $\xi \sim Beta(\lambda)$,
   - *3.2 Update probability measures,* $\mu = N(t_m|\xi, \Sigma)$ *and* $\overline{\boldsymbol{\pi}} = (\boldsymbol{\pi}^{-1}, \mu\pi_k)$, *where* $\boldsymbol{\pi} = (\boldsymbol{\pi}^{-1}, \pi_k)$
   - *3.3 Draw an infinite multinomial* θ_m $\theta_m|\alpha_0, \overline{\boldsymbol{\pi}} \sim Dirichlet(\alpha_0\overline{\boldsymbol{\pi}})$
   - *3.4 For each word n in the document,*
     - *i. Draw a topic* $z_{mn} \sim Multinomial(\theta_m)$
     - *ii. Draw a word* $w_{mn}|z_{mn} \sim Multinomial(\phi_{z_{mn}})$
     - *iii. Draw a time* $t_{mn}|z_{mn} \sim Beta(\psi_{z_{mn}})$

In order to implement the model, we use Gibbs sampling as the inference algorithm. Similar to infinite LDA, this model employs Dirichlet as the base distribution, where $\boldsymbol{\pi} \sim Dirichlet(\gamma/K)$. Note that, we acquire the hyper-parameter via the topics global distribution over time. When updating the measures, we set $\overline{\pi_k} = \xi\pi_k$, where $\overline{\pi} = (\boldsymbol{\pi}^{-1}, \overline{\pi}_k)$, in which time prior is drawn from $\xi \sim Beta(\lambda)$. In this process, we consider K as an infinite variable.

If there exists no time prior to control the probability of generating new topic by DP, the new topics may be distributed averagely over time. But it does not accord with the real topics distribution. With the iteration accumulates, the probability of the topic assignment of every word will be manipulated by time factor gradually. For the Beta distribution will be sharper after updating the posterior distribution. Hence, we expect more topics generated in the timestamp where they congregate in reality and the beta distribution of each topic will be a little smoother relatively. So that we would have a balance between content relevance and time information.

Sampling z. Since the Stick-breaking representation models the topic distribution over terms by Dirichlet distribution which is the same as LDA. The

conditional probability is:

$$p(z_{mn}|\bullet) \propto (n_{m,z_{mn}} + \alpha\overline{\pi}_{z_{mn}}) \cdot \frac{(1-t_{mn})^{\psi_{z_{mn}1}-1} t_{mn}^{\psi_{z_{mn}2}-1}}{B(\psi_{z_{mn}1}, \psi_{z_{mn}2})} \cdot \frac{n_{z_{mn}w_{mn}} + \beta_{w_{mn}} - 1}{\Sigma_{i=1}^{V}(n_{z_{mn}i} + \beta_i) - 1} \quad (7)$$

As there will be new topic being generated, the measures $\boldsymbol{\pi}$ should be updated. Thus, its posterior probability is:

$$\boldsymbol{\pi} \sim Dirichlet(u_1, u_2, ..., u_{k-1}, \gamma) \quad (8)$$

where the u_j is the sum number of words assigned to the jth topic in all documents, and parameter γ manipulate the probability of generating a new topic. The $n_{m,z_{mn}}$ represents the number of words assigned to topic z_{mn} in document m. Besides, the parameters of Beta distribution associated to every topic illustrate the spikes of the trends, and they will be updated as:

$$\psi_{z1} = \bar{t}_z(\frac{\bar{t}_z(1-\bar{t}_z)}{s_z^2} - 1)$$
$$\psi_{z2} = (1-\bar{t}_z)(\frac{\bar{t}_z(1-\bar{t}_z)}{s_z^2} - 1) \quad (9)$$

3.3 Superposition Topic Model

In HDP, topic number is determined by the corpus itself and the hyper-parameters associated with Dirichlet process. Scilicet it finds a most suitable number of the components on the basis of their internal aggregation and discrimination between each other. A new component will be generated when it seems that no component fits the preceding ones. As it mentioned above, the generation of new component via HDP is limited in specific time period. It means that the whole process of clustering is solely based on content relevance. Hence, we present a superposition topic model

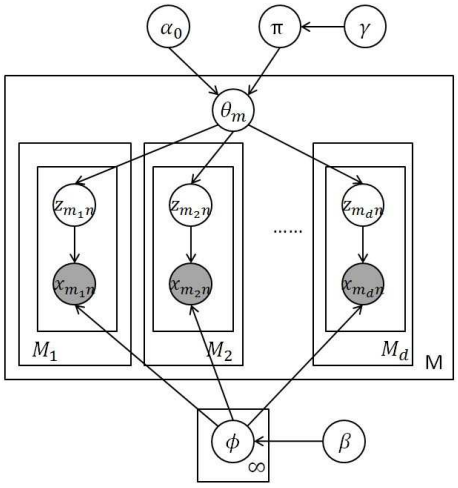


Figure 4: Superposition Topic Model.

(STM) to make use of the dynamic feature of HDP in another perspective.

In the beginning, we discretize the corpus by time beforehand. The STM discovers the topics on the initial part of the corpus without setting number of topics normally. After achieving the initial distributions, the STM proceeds on the following part of corpus, and it is more likely to generate new topics by HDP due to the focal point of documents differs as the time elapses. Likewise, if the contents of new articles are nearly the same as the previous, the topics will retain their old number. From the perspective of Chinese restaurant process, a group of new comers would rather hold together in a new table than join the previous tables with unfamiliar dishes. Hence, every new part of corpus is processed upon the preceding topic distribution and the STM is updated simultaneously. When we employ the STM on news dataset, the progressively generated topics will unfold the diversion of public provenance correspondingly. Its graphic representation is shown in Figure 4, and the generative process is described as follows:

1. Draw an infinite dimension multinomial $\boldsymbol{\pi}$, $\boldsymbol{\pi}|\gamma \sim GEM(\gamma)$
2. For each toic z, draw a multinomial ϕ_z, $\phi_{z_{mn}}|\beta \sim Dirichlet(\beta)$
3. For each part d of the corpus,
   - 3.1 If d $\neq$ 1, update the measures $\boldsymbol{\pi}$, $\boldsymbol{\pi} \sim Dirichlet(u_1, u_2, ..., u_{k-1}, \gamma)$
   - 3.2 Draw an infinite multinomial θ_m $\theta_m|\alpha_0, \boldsymbol{\pi} \sim Dirichlet(\alpha_0\boldsymbol{\pi})$,
   - 3.3 For each word n in the document,
     - i. Draw a topic $z_{mn} \sim Multinomial(\theta_m)$
     - ii. Draw a word $w_{mn}|z_{mn} \sim Multinomial(\phi_{z_{mn}})$

Since the global measures $\boldsymbol{\pi}$ is updated by time order, the content relevance between preceding documents and new arrival ones are the most significant inducement of generating a new topic. Hence, it is no necessary for us to get entangled in subjoining temporal information while increasing the complexity of model structure, which has been explained by Occam's razor primely. The HDP will automatically determine the sampled term to join the previous topics or to be a new one. Apparently, the judgements are effected by time to some big extent.

Sampling z. The conditional probability when sampling z can be achieved via:

$$p(z_{mn}|\bullet) \propto \cdot(n_{m,z_{mn}} + \alpha_0\pi_{z_{mn}}) \cdot \frac{n_{z_{mn}w_{mn}} + \beta_{w_{mn}} - 1}{\Sigma_{i=1}^{V}(n_{z_{mn}i} + \beta_i) - 1} \quad (10)$$

which is similar to traditional LDA. The updating step of hyper-parameters is vital in the gradual inference

Table 1: A comparison of topic distribution.

Original Topic	healthcare	bill	insurance	health	reform	coverage	americans	option	applause	committee	finance	secretary
	0.0480	0.0452	0.0396	0.0387	0.0347	0.0179	0.0168	0.0128	0.0088	0.0078	0.0072	0.0070
Wiki Topic	**Healthcare**	**Bill**	**Insurance**	**Reform**	**Health**	**Coverage**	**Americans**	**Option**	**Debate**	**Secretary**	**Applause**	**House**
	0.0607	0.0568	0.0476	0.0439	0.0423	0.0205	0.0203	0.0155	0.0111	0.0104	0.0101	0.0101

process. Even though all the topics in the corpus are still exchangeable, the words in preceding documents will not be sampled in the following inference process any more. Documents in every part of the corpus are only sampled in the unique sub-corpus which belongs to a specific epoch. The previous sampled words in other epoch will no longer be associated to another topic in case of topic drifting. Hence, the global cluster weighs will be progressively updated and determine the topic assignment of every term in the subsequent documents.

3.3.1 Wikipedia Semantic Knowledge

Wikipedia concepts are commonly employed to overcome the drawback of bag of words(BOW) in text analysis. However, in this paper, we exploit the Wikipedia concepts as an approach to extract entities from specific document due to its ontology property. Besides, we also extract capitalized words for auxiliary, as there exist a minority of abbreviations and human names which Wikipedia is yet incapable to interpret. Hence, we integrate those words with Wikipedia concepts as our entity volume.

For the sake of briefness, we solely extract the part of document in the graphic model representation so that the approach becomes much easier to comprehend and expand. In this model, we employ two variables to model the entities when discovering topics. e_{mc} and $\bar{z}_{mc}$ represent the cth entity in document m

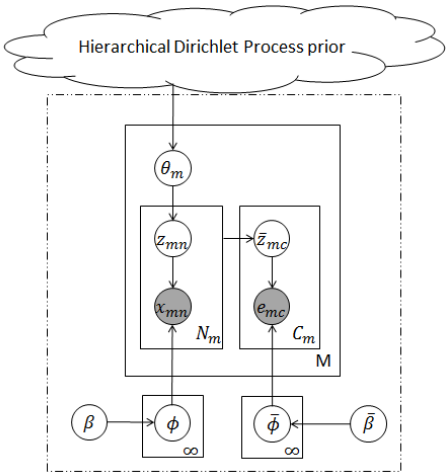


Figure 5: Graphic Model of Wikipedia Semantic Approach from Documental Perspective.

and the assigned topic associated with the entity respectively. According to the graphic model in Figure 5, entity e is an observed variable which depends on assigned topic $\bar{z}_{mc}$ and its topic distribution parameter $\bar{\phi}$. Besides, $\bar{z}_{mc}$ depends on the topic assignment of the word in the document m. The generative process is described as follows:

1. Draw Hierarchical Dirichlet Process prior
2. For each topic z, draw a multinomial ϕ_z, $\phi_{z_{mn}}|\beta \sim Dirichlet(\beta)$
3. For each topic $\bar{z}$, draw a multinomial $\phi_{\bar{z}}$, $\phi_{\bar{z}_{mc}}|\beta \sim Dirichlet(\bar{\beta})$
4. For each document m,
   - 4.1 Draw an infinite multinomial θ_m $\theta_m|\alpha_0, \boldsymbol{\pi} \sim Dirichlet(\alpha_0 \boldsymbol{\pi})$,
   - 4.2 For each word n in the document,
     - i. Draw a topic $z_{mn} \sim Multinomial(\theta_m)$
     - ii. Draw a word $w_{mn}|z_{mn} \sim Multinomial(\phi_{z_{mn}})$
   - 4.3 For each entity c in the document,
     - i. Draw a topic $\bar{z}_{mc}$, $\bar{z}_{mc}|\boldsymbol{Z_m}, N_m \sim Multinomial(\boldsymbol{Z_m}/N_m)$
     - ii. Draw a entity e_{mc}, $e_{mc}|\bar{z}_{mc} \sim Multinomial(\bar{\phi}_{\bar{z}_{mn}})$

Theoretically, after the sampling process, the topic distribution over terms is different from the distribution over entities. And they may differ in the number of topics because of the dynamic property of infinite model. However, after Gibbs sampling process, the representative meanings of them turn out to be extraordinary similar. Moreover, topics over entities achieve better readability and lower perplexity (as illustrated in Table 1). When employed in the scenario of trend tracking, the Wikipedia semantic based approach discovers more trend-specific topical entities, while the increased complexity of inference algorithm is limited in a linear magnitude.

Sampling $\bar{z}$. The z sampling process remains the same as the original form (7) or (10). While, the conditional probability of $\bar{z}$ is:

$$p(\bar{z}_{mn}|\bullet) \propto \cdot \frac{S_{m,\bar{z}_{mc}}}{N_m} \cdot \frac{n_{\bar{z}_{mc}e_{mc}} + \bar{\beta}_{e_{mc}} - 1}{\Sigma_{i=1}^{\bar{V}}(\bar{n}_{\bar{z}_{mc}i} + \bar{\beta}_i) - 1} \qquad (11)$$

Where $S_{m,z_{mc}}$ denotes the sum of words in document m which have been assigned to topic z_{mc} and $S_{m,z_{mc}}/N_m$ represents the prior of generating topic $\bar{z}_{mc}$ in document m and $\bar{V}$ denotes the volume of entities.

4 EXPERIMENT

4.1 BBC News

In this section, we present the discovered topics and their evolutionary processes on 3500 BBC news which is shown in Table 2. We also follow the difference between ITTM and STM with interest. In the experiment, ITTM discovered 90 topics by 1000 iterations, while STM discovered 80 topics by 2400 iterations (200 iterations in every epoch-specified sub-corpus). For the purpose of interpreting model effectiveness, we analyzed the raw corpus by key word matching based on discovered topics. Then we can evaluate the matching degree between tracked trend and reality.

Table 2: Details of BBC news corpus.

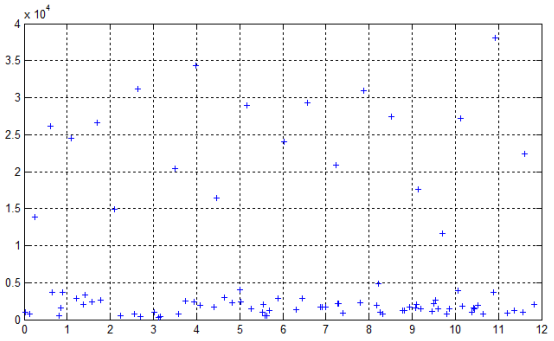
Month	Jan	Feb	Mar	Apr	May	Jun	Jul
Articles	309	284	279	285	279	255	309
	Aug	Sep	Oct	Nov	Dec		
	273	318	260	310	339		

4.1.1 ITTM

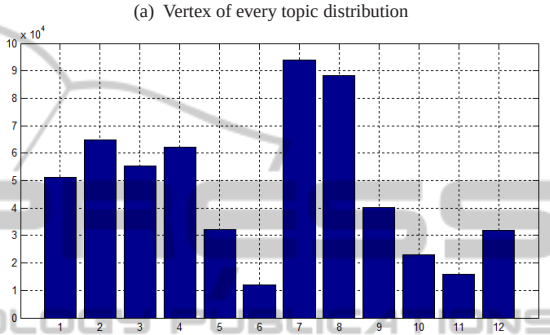
As illustrated in Figure 6(a), there exist about 20 topics with higher heat score while the others stay in a low level respectively. Most of the 20 topics are regarding politics or society, such as *Obama, Military* and *Government*. It illustrates the vane of news trend is focused on politics. In Figure 6(b), we find that the general heat score rises sharply from June to July and August. After looking back to the raw news articles, we are aware of several booming news happened around July and August, such as *Helicopter crash, Michael Jackson's death, Dugard kidnapping case* and so on. But unfortunately, ITTM didn't organize them as single topics, because these topics are more likely to be assigned to the topics with wide coverage like Topic Society or Topic Criminal.

4.1.2 STM

STM discovered 80 topics in the experiment and we recorded the result in every sub-corpus which is shown in Table 3. In the first month, STM generates 34 topics, while the number rises gradually to the end of 80. Since the global probability measure θ is shared in every sub-corpus, we could extract the evolutionary process by tracing the transformation of topic distribution over terms which is represented by ϕ. By calculating the KL divergence between every two topics, we achieve an evolutionary process(shown in Table 4



(a) Vertex of every topic distribution



(b) General topic heat assigned to every month

Figure 6: Topics of ITTM in 2009.

partly). We conclude all the topics into 3 categories, including stable topics, new topics and vanished topics. In Table 4, the topics with grey ground colour change slightly during the whole lifetime and their KL divergence remain below 1.0, so we name them stable topics. Correspondingly, we call the generated topics by STM of every epoch as new topics. Moreover, when an old topic vanishes, a new topic will also be generated concomitantly. The bolded numbers which exceed 1.0 in Table 4 divide the topic into 2 parts, hence the latter ones are taken into account as new topics too, and the formers are perceived vanished topics. Table 5 illustrates the evolution of topics from vanished ones to new ones.

As a whole, the stable topics, new topics and vanished topics contain 50, 49 and 19 respectively. In order to exhibit the evolutionary processes of topics, we present 23 topics in two parts with their heat score in Figure 10. Most topics of Figure 7(a) are stable ones, while in Figure 7(b) most of them are new booming topics. The peaks of broken lines show the trend primely. It is easy for us to demonstrate the causation by big events in 2009, for instance, *"Flu pandemic"* in April, *"Enough bomb-grade uranium of Iran"* in April, *"Helicopter crash of Maryland"* in July, *"Jaycee Lee Dugard abduction case"* in August and *"Fort Hood shootings at US army base"* in November. Meanwhile, the stable topics such as Topic *Politics, Guantanamo, economy, criminal* and

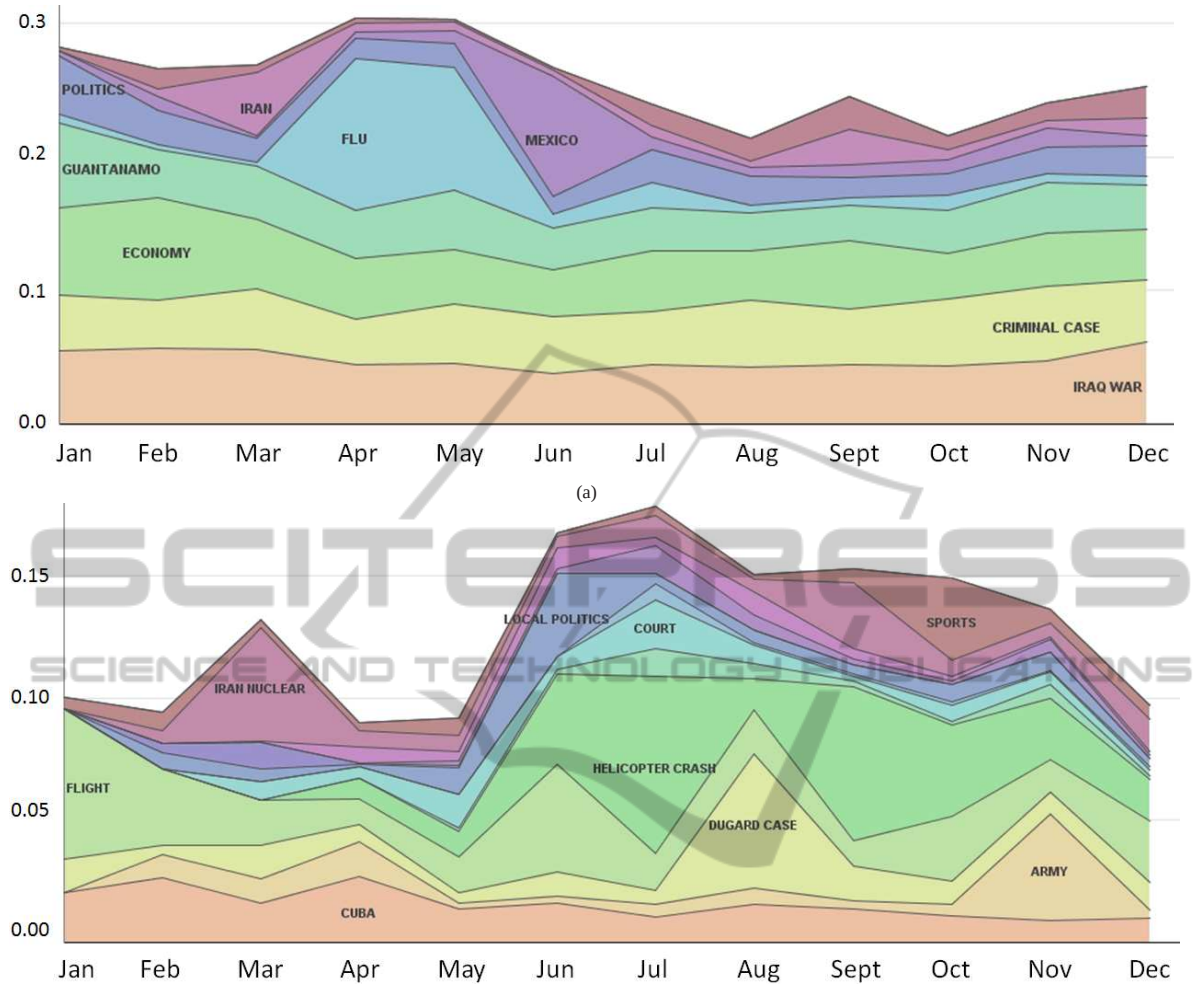


Figure 7: Topic trends of STM.

Iraq war receive persistent public attention by people.

Afterwards, we employed the Wikipedia semantic based approach on STM and achieved better readability and distinctly organized topics. Table 3 shows a significant decline in perplexity and Table 1 presents more topic-representative entities with higher probability, which demonstrate that it contributes to tracking the event to some extent.

4.2 Trend Tracking Analysis

For the purpose of better exhibiting trend tracking result, we draw a curve via key words matching which we consider generally reflects the real trend of each topics. Hence, we obtain Figure 8, including four topics, *'Flu'*, *'Healthcare'*, *'Gay'* and *'Train'*. As it is shown in the figure, ITTM matches the spikes of heat trends precisely, but is incapable to simulate multi-spikes trend of the reality. For instance, in Figure 8(c), the curve drops from June and fails to match the peak in September. Relatively, STM simulates the real trend primely, even though the multi-spikes and a bit fluctuation somewhere in the whole year.

In general, ITTM generates a series of smoothing curves to fit the real trends and extracts the spikes of every topic distribution over time when discovering topics. Nevertheless, STM simulates the trends via topic distributions transformation, for the topics are dominated by global probability measures. Even though these two approaches are based on different assumptions, both of them generally model the whole evolutionary processes of topics.

4.3 Model Effectiveness

On the basis of the experiments above, these findings suggest the models are capable of tracking trends and

Table 3: Experimental results of STM.

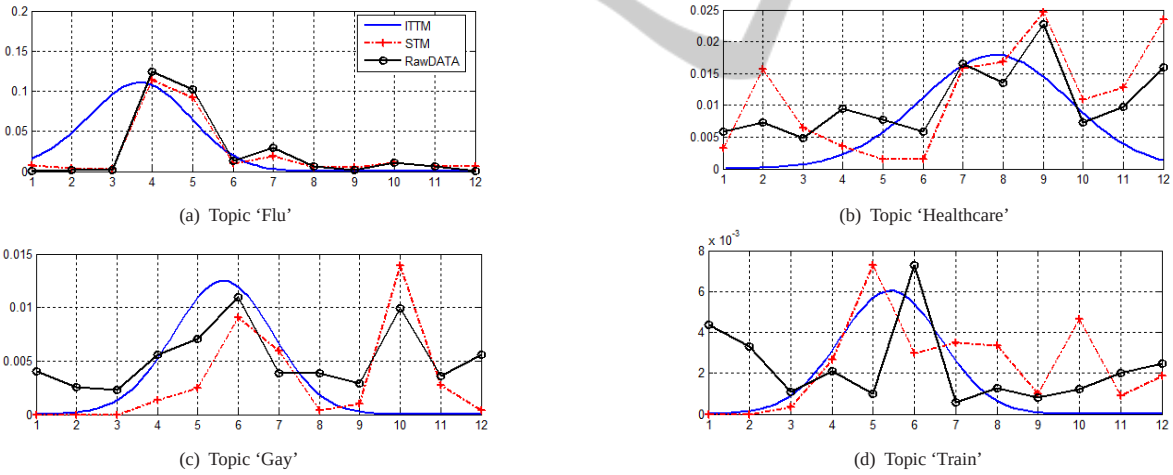
Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
Sum topics	34	48	58	63	67	69	72	74	78	80	80	80
Generated New Topics	34	14	10	7	4	2	3	2	4	2	0	0
Perplexity over Terms	2817	2744	3132	2922	3172	2892	3358	3321	3427	3719	3676	3723
Perplexity over Entities	1465	1668	2102	1910	2100	1899	2139	2157	2187	2414	2368	2342

Table 4: KL Divergence between one topic and its preceding one.

No	Topic	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
Topic2	Politics	-	0.170	0.056	0.036	0.030	0.027	0.043	0.027	0.012	0.008	0.013	0.013
Topic4	Guantanamo	-	0.119	0.071	0.030	0.070	0.016	0.017	0.013	0.012	0.009	0.019	0.010
Topic7	Weather	-	0.217	**1.027**	0.341	0.058	0.276	0.076	**1.994**	0.081	0.017	0.033	0.006
Topic11	Economy	-	0.258	0.056	0.025	0.018	0.011	0.015	0.008	0.012	0.005	0.007	0.006
Topic12	Obama	-	0.058	0.021	0.014	0.008	0.004	0.004	0.003	0.004	0.003	0.003	0.003
Topic13	Criminal	-	0.275	0.147	0.047	0.052	0.021	0.019	0.018	0.015	0.013	0.014	0.008
Topic21	Flu	-	0.631	0.259	**4.143**	0.401	0.008	0.025	0.003	0.005	0.007	0.006	0.005
Topic29	Healthcare	-	**2.998**	0.373	0.072	0.043	0.039	**1.008**	0.227	0.352	0.048	0.094	0.213
Topic36	Army	-	-	**1.338**	**1.585**	0.027	0.061	0.250	0.155	0.061	0.066	**1.890**	0.015
Topic43	Local Politics	-	-	0.777	0.062	**1.345**	**2.559**	0.040	0.045	0.041	0.060	0.157	0.037
Topic44	Neoconservatism	-	-	**1.993**	0.055	0.154	0.001	**1.060**	0.041	0.031	0.116	0.014	0.122
Topic51	Court	-	-	-	0.640	**2.025**	0.215	0.932	0.062	0.039	0.083	0.049	0.014
Topic56	Train	-	-	-	**1.120**	0.427	**2.431**	0.280	0.030	0.074	0.033	0.123	0.013
Topic60	UN	-	-	-	-	**1.357**	**2.755**	**1.346**	0.027	0.451	0.050	0.040	0.018
Topic63	Helicopter Crash	-	-	-	-	-	0.475	**2.010**	0.722	0.264	0.099	0.435	0.074

Table 5: Two examples of topic diversion.

Topic 56th	May	'Train'	British Trains Rail Travel Gordon Services Brown Sexual Peruvian Gdp Runs Minister Sort Position
	Jun	'Accident'	Bermuda Uighurs Train British China Trains Palau London Accident Foreign Government Four
Topic 7th	Jul	'Weather'	Weather Service Sheriff Myers Project Died Police County Storm Brother Snow Ms Couple
	Aug	'Criminal'	Garrido Ms Weather Dugard Police County Service Sheriff Project Myers Jaycee Phillip Probyn



(a) Topic 'Flu'

(b) Topic 'Healthcare'

(c) Topic 'Gay'

(d) Topic 'Train'

Figure 8: Evolutionary processes of topics simulated by ITTM and STM.

Table 6: Topic Perplexity.

Num.of Articles	100	500	1000	2000	3500
ITTM	1493	2852	4288	4454	4514
STM	1435	2138	2640	3068	3551
STM&Wiki	**1023**	**1336**	**1629**	**1779**	**2249**
infiniteLDA	1432	2008	2386	2718	3098

Table 7: Experiment Results on Jan. 2010 corpus.

Model	ITTM	STM	STM&Wiki	infiniteLDA
Sum topics	63	85	85	66
Perplexity	2365	2385	1461	2295

receive a series of desirable results. Likewise, we did further experiment on different magnitude corpus to reveal the effectiveness of each model.

From the data in Table 6, we employ ITTM, STM, STM&Wiki and infinite LDA on those corpuses. The results indicate that STM&Wiki obtain the best performance, while the perplexity of ITTM is slightly bigger than others. Furthermore, we prepared a corpus of Jan. 2010 which contains 344 articles for trend prediction. Table 7 indicates the perplexity compari-

Table 8: Trend Prediction.

Model		Number of Articles in each Trend												
ITTM	Articles	6	11	17	10	15	13	19	19	12	**129**	7	6	9	8	7
	Related	5	4	8	5	8	8	12	3	8	**60**	4	4	3	5	4
STM	Articles	7	9	16	6	85	8	15	**96**	17	13	6	6			
	Related	3	5	6	3	48	5	6	**42**	10	9	4	3			
iLDA	Articles	19	7	9	8	6	7	11	**183**	44	18					
	Related	8	4	3	3	3	3	5	**83**	18	11					

son on topic inference between these models.

Then we organized clustered articles by the models, and made a manual evaluation (based on article title and news description) as shown in Table 8. We got the sum precision about trend prediction of these models. Each of ITTM, STM and infinite LDA is 0.4896, 0.5070 and 0.4519. Interestingly, we find some trends contain much more articles than the other ones. The reason is that in the middle of Jan. 2010, a powerful earthquake rocks Haiti which triggered a series of news reports on this disaster. Most articles in these trends are concerning this event. After all, both of ITTM and STM can predict a real world trend successfully, even though on the booming event like "Haiti earthquake".

5 CONCLUSIONS AND FUTURE WORK

In this paper, we present two approaches incorporating HDP and temporal information on real-world task without Markov assumption. Meanwhile, a Wikipedia semantic based approach has been exploited to improve the results of topic modelling. Namely, the models hold the complexity in a low level with succinct graphic representation. The experimental results indicate the capability of tracking trend from news media. As a significant finding, the ITTM simulates the peak of event trend precisely but fails to handle the multi-spikes situation. While the STM is capable of tracking the trends with fluctuations and discovering new topics, stable topics and vanished topics. Because of the flexibility and no number limitation of topics, the models can be easily extended to other scenarios. Our future work might focus on tracking the user interest by incorporating propagation algorithms based on proposed models. The combination of infinite topic modelling and location factor is also under our consideration.

REFERENCES

Ahmed, A. and Xing, E. P. (2010). Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. *In UAI '10*.

AlSumait, L., Barbara, D., and Domeniconi, C. (2008). On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. *In ICDM '08*, pages 3–12.

Balasubramanyan, R., Cohen, W. W., and Hurst, M. (2009). Modeling corpora of timestamped documents using semisupervised nonparametric topic models. In *NIPS*.

Blei, D., Ng, A., Jordan, M., and Lafferty, J. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(993-1022).

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In *ICML*.

Ferguson, T. (1973). Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230.

Heinrich, G. (2011). "infinite lda"-implementing the hdp with minimum code complexity. Tecnical Note.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *SIGIR*.

Hong, L., Yin, D., Guo, J., and Davison, B. D. (2011). Tracking trends: Incorporating term volume into temporal topic models. In *KDD*.

Kataria, S. S., Kumar, K. S., Rastogi, R., Sen, P., and Sengamedu, S. H. (2011). Entity disambiguation with hierarchical topic models. In *KDD*.

Landauer, T. K.and Dumais, S. T. (1997). A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(211-240).

Lau, J. H., Grieser, K., Newman, D., and Baldwin, T. (2011). Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1536–1545.

Newman, D., Chemudugunta, C., and Smyth, P. (2006). Statistical entitytopic models. In *KDD*.

Ni, X., Sun, J.-T., Hu, J., and Chen, Z. (2009). Mining multilingual topics from wikipedia. In *WWW*.

Ren, L., Dunson, D. B., and Carin, L. (2008). The dynamic hierarchical dirichlet process. In *ICML*.

Sudderth, E. B. (2006). Graphical models for visual object recognition and tracking. *Doctoral Thesis, Massachusetts Institute of Technology*.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(1566-1581).

Wang, C., Blei, D. M., and Heckerman, D. (2008). Continuous time dynamic topic models. *In UAI '08*, pages 579–586.

XueruiWang and McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In *KDD*.

Zhang, J., Song, Y., Zhang, C., and Liu, S. (2010). Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *KDD*.