TUT: A Statistical Model for Detecting Trends, Topics and User Interests in Social Media

Xuning Tang
College of Information Science and Technology
Drexel University
Philadelphia, PA, USA
xt24@drexel.edu

Christopher C. Yang
College of Information Science and Technology
Drexel University
Philadelphia, PA, USA
chris.yang@drexel.edu

ABSTRACT
The rapid development of online social media sites is accompanied by the generation of tremendous web contents. Web users are shifting from data consumers to data producers. As a result, topic detection and tracking without taking users' interests into account is not enough. This paper presents a statistical model that can detect interpretable trends and topics from document streams, where each trend (short for trending story) corresponds to a series of continuing events or a storyline. A topic is represented by a cluster of words frequently co-occurred. A trend can contain multiple topics and a topic can be shared by different trends. In addition, by leveraging a Recurrent Chinese Restaurant Process (RCRP), the number of trends in our model can be determined automatically without human intervention, so that our model can better generalize to unseen data. Furthermore, our proposed model incorporates user interest to fully simulate the generation process of web contents, which offers the opportunity for personalized recommendation in online social media. Experiments on three different datasets indicated that our proposed model can capture meaningful topics and trends, monitor rise and fall of detected trends, outperform baseline approach in terms of perplexity on held-out dataset, and improve the result of user participation prediction by leveraging users' interests to different trends.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications – Data Mining

General Terms
Algorithms, Experimentation

Keywords
Topic, Trend, User Interest, Evolution, Modeling.

1. INTRODUCTION
Different from the age of Web 1.0, nowadays web contents are mainly generated by web users. To catch the pulse of a rapidly changing world, it is critical to model the evolution of topics and users' interests over time in a streaming fashion. Therefore, given a document corpus, the goal of this work is proposing a model to discover trending stories and their representative vocabularies, to detect topics and their evolutions, to classify each individual document to a trending story, and to understand web users' interests toward the identified trending stories. This will bring great opportunities for news/blog recommendation, personalized search, and marketing.

From a statistical point of view, a topic is a cluster of words frequently co-occurred. For instance, news articles about "European sovereign debt crisis" might contain topics such as finance, trading, election and foreign relationship. Similarly, news articles about "Withdrawal of U.S. military force" might involve topics such as election, foreign relationship, military and homeland security. In this paper, we consider "European sovereign debt crisis" and "Withdrawal of U.S. military force" in the examples above as two **trending stories** (**trend for short**). Different trending stories can share one or more topics. Specifically, a trend can be a series of continuous events or a storyline, which can rise or fall over time. In addition, as it was evidenced in many online social media sites, users' interests play an important role to the generation of web contents. Users interested in a specific trend have more motivation to contribute their opinions, observations or thoughts to online social media sites. Therefore, we consider user interests in this paper as their interests toward trends instead of topics. To illustrate trend, topic and user interest, figure 1 below shows three different trends (trend I, trend II and trend III). For each trend, we plot its volume (y-axis) over 4 consecutive time intervals (x-axis). Each trend may contain different topics indicated by different colors. A topic can be shared by different trends. In addition, in different time interval, the portion of a given topic inside a trend can be different too. Last but not least, for each trend, users are ranked according to their interests to this trend within each time interval.

Most of the existing works regarding topic detection and tracking employed either generative approaches or clustering-based techniques. These methods retrieved topics as bags of words and captured the potential relationships between topics according to their word-based similarities, e.g. using KL-Divergence to quantify topic similarities when each topic was represented by a distribution of words [9, 11]. However, the topic evolution result obtained by these approaches is hard to interpret, as each ongoing trend might contain multiple topics and each topic can be shared by many different trends. For instance, 2012 NCAA tournament and 2012 UEFA Champions League are two completely different trends, although the news articles about NCAA tournament and the news articles about UEFA Champions League share several common topics such as Sports, Commercial Promotion and Security. By merely knowing the rise or fall of a specific topic, for example security topic, without knowing which trend the topic belongs to, readers will find it difficult to capture the major trending stories. Some most recent works tried to address this

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'12, October 29– November 2, 2012, Maui, HI, USA.
Copyright 2010 ACM 978-1-4503-1156-4/12/10…$15.00.

problem by introducing the concept of "trend"[9] (or "story" [1]). Generally speaking, Kawamae introduced a trend class variable into his Trend Analysis Model (TAM), where each document belongs to a trend and each trend has a probability distribution over topics and words. Unfortunately, the number of trends in TAM is fixed over time and needs to be predefined, which might be difficult in real-world application. Furthermore, to our best knowledge, none of the existing topic detection and tracking works has incorporated dynamic user interests into their models. Different from the previous works which extracted topics from news articles or scientific literatures, our focus is to detect dynamic trends and topics from user-generated contents in online social media. As a result, user interests must be taken into consideration. In addition, incorporating user interests into topic model might help the model to better generalize to unseen data.

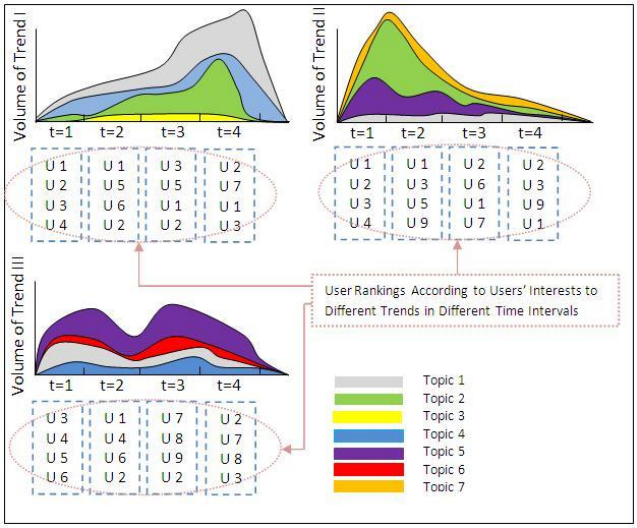


Figure 1. Illustration Example of Concepts about Trend, Topic and User Interest

In this paper, we propose a new probabilistic generative model to simulate the generation process of both web contents and users' participations in a unified framework. In our model, we use a latent variable to represent the trend label of each document which has benefits including: a) a document corpus can be well organized and navigated easily as documents within a corpus are classified according to their trend label; b) a detected trend can roughly represent a storyline or a series of continuous events which increases its interpretability; c) the word distribution of a given trend shows the signature terms of this trend which, to some extent, can be used to label a trend; d) each word can be generated by a background topic, a trend-specific background topic or one of the general topics shared by the whole corpus. In addition, our proposed model can determine the appropriate number of trends automatically in a document corpus by incorporating a RCRP. Last but not least, compared to existing works, another distinct feature of our model is that users' participations are modeled simultaneously during the generation process. It provides the opportunities of offering applications such as news recommendation, personalized search etc.

In the next section, we review some existing works about topic detection and tracking. We pay particular attention to two models, RCRP and TAM, which motivate our model and serve as baseline in the experiment section.

2. LITERATURE REVIEW

In the past decade, topic modeling has attracted many research efforts from both academia and industry. A topic consists of a cluster of words frequently co-occurred. Given a corpus, a topic model can distinguish words with different semantic meanings and extract hidden topics. Hofmann [8] proposed the PLSI model given the assumption that each document contains different topics and each word is generated by one of these topics. Blei et al. [4] proposed the LDA model which used a Dirichlet prior to solve the over-fitting problem which makes the model more flexible. As further extensions, Rosen-Zvi et al. proposed the author-topic model [13], and McCallum et al. proposed the author-recipient-topic-model [10]. Recently, Liu et al. [18] developed a Bayesian hierarchical approach to perform topic modeling and author community discovery in a unified framework. However, all these works only investigated static topics or users without considering their evolutions.

Beyond static topic modeling, other researchers of Topic Detection and Tracking (TDT) extend topic modeling to take into account of their evolutions. Some researchers studied TDT problem by using generative models. In Mei and Zhai's work [11], text streams were partitioned into intervals. They employed a probabilistic mixture model to infer topics within each interval and discovered their evolutionary transitions using KL-divergence. Wang and McCallum [17] proposed a Topic-Over-Time model (TOT) to generate topics. Each generated topic was associated with a continuous distribution over time. AlSumait et al presented an online version LDA model which automatically captures the thematic patterns and identifies emerging topics of text stream[3]. Similarly, Blei and Lafferty [5] designed a Dynamic Topic Model (DTM) by using Gaussian time series and logistic normal topic proportion models. Qi et al proposed to make use of citations to extract evolutionary relationships between topics [6]. However, none of these TDT models takes user information into consideration which is critical for modeling user-generated contents in online communities. Lin et al. [7] proposed to model topic evolution along with the burstiness of user interest. However, they modeled topics using a PLSI-like approach which suffered from the over-fitting problem.

Other than using generative models, some other research works rely on clustering approach to detect evolving topics. Each cluster corresponds to a set of documents with a focused theme. Morinaga and Yamanishi [12] treated documents as an incoming stream and then used a finite mixture model to cluster documents by maximizing their posterior probability. They also applied the theory of dynamic model selection to detect emerging topics indicated by changes of the mixture model. Schult and Spiliopoulou [14] proposed to use clustering approach to identify emerging and persistent taxonomy or ontology for documents. Shahaf et al introduced a novel method and a set of measurements to construct an information map, namely as metro map, to capture the connections of articles and provide structured summaries of information[15]. Recently, Ahmed and Xing extended the Chinese Restaurant Process to Recurrent Chinese Restaurant Process which can be used for evolutionary clustering for document corpus [2]. Concretely, the RCRP is formally defined as:

$$\theta_{t,i}|\{\theta_{t-1,.}\},\theta_{t,1:i-1},\alpha,G_0 \sim \frac{1}{N^{(t-1)}+i-1+\alpha}\times$$
$$\left[\sum_{k\in I_{t-1}}\left(n_{k,t-1}+n_{k,t}^{(i)}\right)\delta(\phi_{k,t})+\sum_{k\in I_t^{(i)}-I_{t-1}}\left(n_{k,t}^{(i)}\right)\delta(\phi_{k,t})+\alpha G_0\right]$$

where $\theta_{t,i}$ is a parameter of a likelihood function $F(\theta_{t,i})$ which generates an observation i at time t, $o_{t,i}$. The equation above can be illustrated by a RCRP metaphor. Assuming that customers entered a restaurant in a given day are not allowed to stay beyond this day, at the end of day $t-1$, the owner of the restaurant records on each table the dish served on this table and the number of customers who shared it. When the i[th] customer $c_i^{(t)}$ enters this restaurant at day t, he can pick a non-empty table k that already has $n_{k,t}^{(i)}$ customers at day t before $c_i^{(t)}$ comes in with probability $\frac{n_{k,t-1}+n_{k,t}^{(i)}}{N^{(t-1)}+i-1+\alpha}$ and use the dish $\phi_{k,t}$. If this table did not exist at day $t-1$, $n_{k,t-1}$ equals to 0. Otherwise, $n_{k,t-1}$ equals to the number of customers who sit on this table at day $t-1$. Alternatively, he can pick an empty table that nobody is sitting on at day t but $n_{k,t-1}$ customers sit on at day $t-1$ with probability $\frac{n_{k,t-1}+n_{i,k}^{(t)}}{N^{(t-1)}+i-1+\alpha}$ and then order a dish $\phi_{k,t}$, where $n_{i,k}^{(t)}$ equals to 0 in this case. Finally, he can pick an empty new table with probability $\frac{\alpha}{N^{(t-1)}+i-1+\alpha}$ and order a new dish based on G_0. We emphasis the RCRP here since it is used as an important component in our proposed model later.

Recently, Ahmed et al. proposed a new model, namely Storyline, which extended the traditional LDA model by incorporating the RCRP [1]. Storyline model combines the power of evolutionary clustering and topic modeling. Therefore it can simultaneously group articles into storylines and identify prevalent topics. The Storyline model has distinctive differences to our proposed model in several aspects: 1) Our model captures user interests toward detected trends while Storyline completely neglect user information; 2) Compared to Storyline, our model uses a switch variable to control topic selection; 3) The generation process of topics in Storyline model and our proposed model are also substantially different. Last but not least, Kawamae proposed a Trend Analysis Model (TAM) which extends the TOT model and introduces a latten trend class variable into each document [9]. As shown in Figure 2, in TAM, each document is assigned to a trend, c, according to a multinomial distribution ψ. Given a trend class c, for each token w_i in this document, a topic is chosen based on a multinomial distribution θ_c. At the same time, a switch variable r_i is sampled. According to the value of r_i, token w_i can be generated based on either a background topic, a trend-related background topic or one of Z topics. By introducing the concept "trend" into the model, some words can be generated directly from the trend class rather than from topics, which helps to model the generation process of the corpus more precisely. Further, relationships between trend and topic are also modeled. However, TAM model has a few limitations as follows: 1) the number of trend is difficult to predefine for practical problems, 2) the model is unsuitable for processing document stream and 3) valuable user information in online social media is completely neglected.

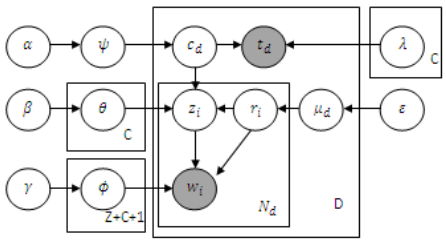


Figure 2. Trend Analysis Model

3. Topic-User-Trend Model (TUT)

In this section, we introduce our Topic-User-Trend model (TUT) to detect evolving topics and trends, and discover users' interests. This model depicts the generation process of web contents and user participations in a unified framework. By estimating this model, we can identify the major trends discussed by users in online social media, learn the topical structure of web contents, and understand users' interests to the detected trends.

Table 1. Notations

E^t, e_d^t	Trend labels of all documents in time interval t, the trend label of document d of time interval t
P, P_u	All users , the user u
Z, z_{d_i}	All topics , the topic of the i^{th} word in document d
W, w_{d_i}	All words , the i^{th} word in document d
X, x_{d_i}	All switch variables , the switch variable of the i^{th} word in document d
α	α is a concentration parameter
$\gamma, \eta, \beta, \iota, \epsilon$	parameters of Dirichlet priors
ϕ_z	the multinomial distribution of words specific to topic z
Ω_e	the multinomial distribution of users specific to trend e
θ_e	the multinomial distribution of topics specific to trend e
φ_e	the multinomial distribution of words specific to trend e
μ_d	the multinomial distribution of switch variables specific to document d
$N^{(t)}$	total number of documents in time interval t
$n_i^{(t)}$, $n_{-d,i}^{(t)}$	number of documents assigned to trend i in time interval t, number of documents assigned to trend i in time interval t except for document d
$K_{d,x=0,1,2}$	number of word tokens assigned to switch x = 0/1/2 in document d
$M_{z,w}$	number of times that word w is assigned to topic z
$N_{e,w}$	number of times that word w is assigned to trend e
$L_{e,z}$	number of times that topic z is assigned to trend e
$C_{e,u}$	number of times that user u is assigned to trend e

3.1 Formulation of Topic-User-Trend Model

Topic-User-Trend model (TUT) is designed based on a real generation process of content terms and users. In online social media, we consider each document (e.g. a thread, a post, or a chain of retweet) as a collaborative contribution by its participants. In addition, it's the trend itself, rather than the topics, that encourages the participation of users. For instance, although threads of NCAA tournament and threads of UEFA Champions League share common topics such as Sports, Commercial Promotion and Security, college basketball fans may participate more in the threads related to 2012 NCAA basketball tournament rather than the threads related to 2012 UEFA Champions League.

In this work, we divide a time axis into several non-overlapping time intervals. Within each time interval, when a new document arrives, we first model the generation process of its trend label by

a RCRP. Once the trend label of this new document is determined, users' interests play an important role to decide whom this new document will attract. In addition, depending on the trend label of this document, we model the generation process of its content terms by a variation of LDA model. Specifically, we assume that the generation of each content term is influenced by one of the three factors: 1) a general background topic, 2) a trend-specific background topic (e.g. 2012 NCAA basketball tournament or 2008 U.S. Presidential Election). and 3) topics shared by the whole corpus (e.g. economy topic, sport topic or military topic). Terms that are topic-unrelated but widely exist in the corpus have a higher chance coming from the general background topic. In addition, each trend may contain some signature terms serving as the trend-specific background terms. At last, similar to LDA model, the entire corpus shares a mixture of |Z| different topics. We adopt a switch variable to control the influence of these three factors in content term generation. It's important to note that the number of topics, |Z|, is predefined and fixed while the number of trends is theoretically countable infinity in our model. Different trends can share common topics. Our proposed model is shown in Figure 3. The meanings of notations in Fig. 3 are listed in Table 1.

As shown in Fig. 3(a), TUT model is a time-dependent model. Each component represents the generation process of documents in a time interval. Fig. 3(b) shows the detail of the TUT model in the t^{th} time interval. Within time interval t, the trend label e_d^t of document d is sampled by a RCRP. If e_d^t is a new trend, a new distribution over users ($\Omega_{e_d^t}$), a new distribution over topics ($\theta_{e_d^t}$) and a new distribution over words ($\varphi_{e_d^t}$) are drawn for this new trend. Otherwise, the existing $\Omega_{e_d^t}$, $\theta_{e_d^t}$ and $\varphi_{e_d^t}$ will be used. Let $|U_d|$ equal to the number of users participated in document d, the user list of d is generated by repeating the sampling process $|U_d|$ times based on a trend-user distribution $\Omega_{e_d^t}$. Each word w_{d_i} is drawn either from the general background topic or the trend-specific background topic or one of |Z| topics shared by the whole corpus. As mentioned, a switch variable x_{d_i} is drawn first from a multinomial distribution for word w_{d_i} to control its generation process. If $x_{d_i} = 0$, w_{d_i} is drawn from the general background topic. If $x_{d_i} = 1$, w_{d_i} is drawn from the trend-specific background topic, $\varphi_{e_d^t}$, which consists of the signature terms of trend e_d^t. If $x_{d_i} = 2$, a topic z_{d_i} is first sampled from the trend-topic distribution $\theta_{e_d^t}$, and then the word w_{d_i} is drawn according to the topic. Overall, the generation process of users and words in the TUT model can be described as follows:

For each time interval t from 1 to T:

1. Draw 1+|Z| multinomial distributions ϕ_z from prior γ, one for each topic (a general background topic and Z topics)
2. Draw |D| multinomial distribution μ_d from prior ϵ, one for each document
3. Draw |E| multinomial distributions Ω_e, θ_e and φ_e from prior η, β and ι respectively, one for each existing trend

For each document d of time interval t:

A. Draw a trend label e_d^t from RCRP
B. If e_d^t is a new trend, Draw multinomial distributions Ω_e, θ_e and φ_e from prior η, β and ι respectively for this new trend
C. Draw $|U_d|$ users from multinomial distribution $\Omega_{e_d^t}$

For each word token w_{d_i} in document d:

Draw switch variable x_{d_i} from multinomial μ_d

if $x_{d_i} = 0$

i. Draw word w_{d_i} from multinomial ϕ_{bg}

else if $x_{d_i} = 1$

i. Draw word w_{d_i} from multinomial $\varphi_{e_d^t}$

else if $x_{d_i} = 2$

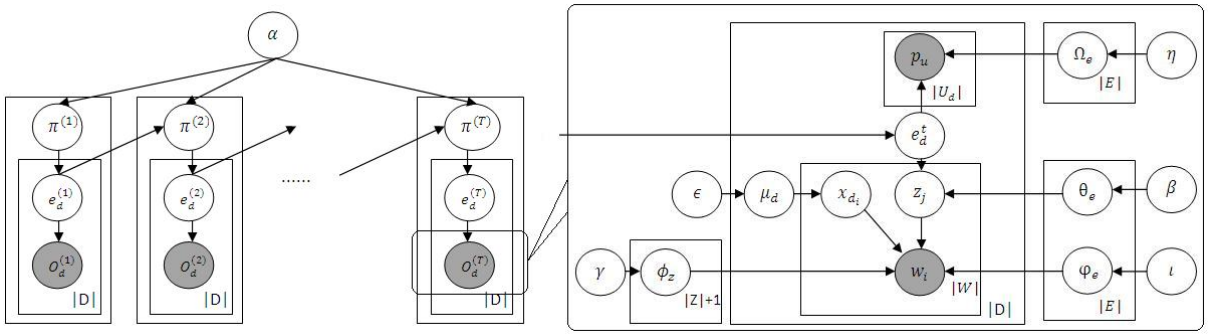
i. Draw topic z_{d_i} from multinomial $\theta_{e_d^t}$

ii. Draw word w_{d_i} from multinomial $\phi_{z_{d_i}}$

3.2 Inference and Learning

Several methods have been developed to estimate the latent variables in a probabilistic graphical model. Among them, Gibbs sampling often yields relatively simple algorithm for high-dimensional data model. In the Gibbs sampling schema, Markov chain is constructed for simulating the generation processes of terms and users. The transition between successive states in the Markov chain is achieved by repeatedly sampling topic and trend for each observed term and user based on its conditional probability. We first provide the joint distribution and then derive conditional probabilities for trend label and topics respectively in section 3.2.1 and 3.2.2.

For simplicity, we use Δ_t to denote all hidden variables of the t[th] time interval and O_t to denote all observations of the t[th] time interval including words and users. Instead of finding a global optimization of $\{\Delta_t\}_{t=1}^T$, given the whole observation sequence $\{O_1, O_2, \ldots, O_T\}$, our goal is to find greedy optimization configuration as did in [16]. Specifically, we assume that we only observe O_{t-1} and O_t. We want to maximize the posterior



(a) (b)

Figure 3. Graphical Model of Topic-User-Trend Model

probability $\Pr(\Delta_t|\Delta^*_{t-1}, O_{t-1}, O_t)$, where Δ^*_{t-1} represents the optimal hidden variables estimated for time interval t-1. The justification is that, instead of observing the whole observation sequence $\{O_1, O_2, \dots, O_T\}$ at the same time, it is more practical to assume that the data comes in a stream fashion and we update the model sequentially.

In the inference process, we need to first calculate the conditional distributions. As shown in the previous subsection, the joint distribution of the documents of time t is computed as:

$$\Pr(E^t, P, Z, W, X|E^{t-1}, \alpha, \gamma, \eta, \beta, \iota, \epsilon)$$
$$= \int\int\int\int\int \Pr(E^t, P, Z, W, X, \phi, \Omega, \theta, \varphi, \mu|E^{t-1}, \alpha, \gamma, \eta, \beta, \iota, \epsilon) d\mathcal{H}$$
$$= \int\int\int\int\int \prod_d^D [\Pr(e_d^t|E^{t-1}, E^t_{-d}, \alpha) \prod_u^{U_d} \Pr\left(P_u\middle|\Omega_{e_d^t}\right)$$
$$\times \prod_i^{N_d} (\Pr(x_{d_i}|\mu_d) \Pr(z_{d_i}|\theta_{e_d^t}) \Pr(w_{d_i}|x_{d_i}, \phi_{z_{d_i}}, \varphi_{e_d^t}))]$$
$$\times \prod_d^D \Pr(\mu_d|\epsilon) \prod_z^Z \Pr(\phi_z|\gamma) \prod_e^{E^t} (\Pr(\Omega_e|\eta) \Pr(\theta_e|\beta) \Pr(\varphi_e|\iota))\, d\mathcal{H}$$

where $d\mathcal{H} = \mathrm{d}\phi d\Omega d\theta d\varphi d\mu$. In the joint distribution above, multinomials $(\phi, \Omega, \theta, \varphi, \mu)$ can be adapted by the conjugate priors $(\gamma, \eta, \beta, \iota\, \epsilon)$ and then integrated out eventually. The conjugate priors can be predefined in this work. Therefore, we only need to estimate e_d^t, z_{di} and x_{di}. The sampling scheme consists of two steps. The first step is to sample the trend label, e_d^t, for a document d of time period d, $\Pr(e_d^t = e| \dots)$. The second step is to sample the switch variable and the topic for each individual word in a document: $\Pr(x_{-di} = 0| \dots), \Pr(x_{-di} = 1|e_d^t = i \dots), \Pr(z_{-di} = k, x_{-di} = 2| \dots)$. Details of the derivation of Gibbs sampling for TUT is given below.

3.2.1 Trend Label Sampling

For each document, we use the chain rule to obtain the conditional distribution to sample its trend label e_d^t. The sampling equation is defined as:

$$\Pr(e_d^t = e| \dots)$$
$$\propto \left[\frac{\prod_u^U \Gamma(\mathrm{C_{e,u}} + \eta_u)}{\prod_u^U \Gamma(\mathrm{C_{e,u\backslash d}} + \eta_u)} \cdot \frac{\Gamma(\sum_{\mathrm{u}}^{\mathrm{U}}(\mathrm{C_{e,u\backslash d}} + \eta_u))}{\Gamma(\sum_{\mathrm{u}}^{\mathrm{U}}(\mathrm{C_{e,u}} + \eta_u))}\right] \times \left[\frac{\prod_z^Z \Gamma(L_{e,z} + \beta_z)}{\prod_z^Z \Gamma(L_{e,z\backslash d} + \beta_z)}\right.$$
$$\left. \cdot \frac{\Gamma(\sum_{\mathrm{z}}^{\mathrm{Z}}(L_{e,z\backslash d} + \beta_z))}{\Gamma(\sum_{\mathrm{z}}^{\mathrm{Z}}(L_{e,z} + \beta_z))}\right] \times \left[\frac{\prod_w^W \Gamma(\mathrm{N_{e,w}} + \iota_w)}{\prod_w^W \Gamma(\mathrm{N_{e,w\backslash d}} + \iota_w)} \cdot \frac{\Gamma(\sum_{\mathrm{w}}^{\mathrm{W}}(\mathrm{N_{e,w\backslash d}} + \iota_w))}{\Gamma(\sum_{\mathrm{w}}^{\mathrm{W}}(\mathrm{N_{e,w}} + \iota_w))}\right]$$
$$\times \begin{cases} \frac{\alpha}{N^{(1)} - 1 + \alpha} & new\ event\ when\ time\ t = 1 \\ \frac{n^{(1)}_{-d,e}}{N^{(1)} - 1 + \alpha} & existing\ event\ when\ time\ t = 1 \\ \frac{\alpha}{N^{(t)} + N^{(t-1)} - 1 + \alpha} & new\ event\ when\ time\ t > 1 \\ \frac{n^{(t)}_{-d,e} + n_e^{(t-1)}}{N^{(t)} + N^{(t-1)} - 1 + \alpha} & existing\ event\ when\ time\ t > 1 \end{cases}$$

where $\mathrm{C_{e,u\backslash d}}$ represents the number of times that user u is assigned to trend e, except in document d. Similarly, $L_{e,z\backslash d}$ represents the number of times that topic z is assigned to trend e, except in document d; $\mathrm{N_{e,w\backslash d}}$ represents the number of times that word w is assigned to trend e, except in document d. The terms within the first square bracket measure the posterior probability of observing users in document d given the trend label $e_d^t = e$. Similarly, the term in the second and third square bracket measures the posterior probability of observing topics and trend-specific background words respectively given $e_d^t = e$. The terms behind the curly bracket denote the probability of assigning document d to trend e by following the RCRP [2]. When time t = 1 which means historical trend information is unavailable, the top half of the term is used ($\frac{\alpha}{N^{(1)}-1+\alpha}$ chance belonging to a new trend and $\frac{n^{(1)}_{-d,e}}{N^{(1)}-1+\alpha}$ chance belonging to an existing trend e) When time t > 1, the second half of the term is used ($\frac{\alpha}{N^{(t)}+N^{(t-1)}-1+\alpha}$ chance belonging to a new trend and $\frac{n^{(t)}_{-d,e}+n_e^{(t-1)}}{N^{(t)}+N^{(t-1)}-1+\alpha}$ chance belonging to an existing trend e)

3.2.2 Switch Variable and Topic Sampling

By using the chain rule again, for each word token, the posterior probability of adding word w_{d_i} in document d to background topic is derived as:

$$\Pr(x_{-di} = 0| \dots)$$
$$\propto \frac{K_{d,0} + \epsilon_0 - 1}{\sum_{\mathrm{x}}^{\mathrm{X}}(\mathrm{K_{d,x}} + \epsilon_x) - 1} \cdot \frac{\mathrm{M_{z=bg,w}} + \gamma_w - 1}{\sum_{\mathrm{w}}^{\mathrm{W}}(\mathrm{M_{z,w}} + \gamma_w) - 1}$$

where the first term measures the probability of having the switch variable equals to 0, and the second term measures the probability of generating w_{d_i} from the background topic.

Similarly, the posterior probability of adding word w_{d_i} in document d to trend i is derived as:

$$\Pr(x_{-di} = 1|e_d^t = i \dots)$$
$$\propto \frac{K_{d,1} + \epsilon_1 - 1}{\sum_{\mathrm{x}}^{\mathrm{X}}(\mathrm{K_{d,x}} + \epsilon_x) - 1} \cdot \frac{\mathrm{N_{e=i,w}} + \iota_w - 1}{\sum_{\mathrm{w}}^{\mathrm{W}}(\mathrm{N_{e,w}} + \iota_w) - 1}$$

where the first term measures the probability of having the switch variable equals to 1, and the second term measures the probability of generating w_{d_i} from the trend-specific background topic.

Similarly, the posterior probability of adding word w_{d_i} in document d to topic k is defined as:

$$\Pr(z_{-di} = k, x_{-di} = 2| \dots)$$
$$\propto \frac{K_{d,2} + \epsilon_2 - 1}{\sum_{\mathrm{x}}^{\mathrm{X}}(\mathrm{K_{d,x}} + \epsilon_x) - 1} \cdot \frac{L_{e,z=k} + \beta_k - 1}{\sum_{\mathrm{z}}^{\mathrm{Z}}(L_{e,z} + \beta_z) - 1} \cdot \frac{\mathrm{M_{z=k,w}} + \gamma_w - 1}{\sum_{\mathrm{w}}^{\mathrm{W}}(\mathrm{M_{z,w}} + \gamma_w) - 1}$$

where the first term measures the probability of having the switch variable equals to 2, the second term measures the probability of selecting topic k in trend e, and the third term measures the probability of generating w_{d_i} from topic k.

3.2.3 Parameter Estimation

Once the sampling processes converge based on the conditional distributions derived in sections 3.2.1 and 3.2.2, we can estimate the five parameters using the following equations:

$$\theta_{e,k} = \frac{L_{e,k} + \beta_k}{\sum_{\mathrm{z}}^{\mathrm{Z}}(L_{e,z} + \beta_z)},\ \phi_{z,w} = \frac{\mathrm{M_{z,w}} + \gamma_w}{\sum_{\mathrm{w}}^{\mathrm{W}}(\mathrm{M_{z,w}} + \gamma_w)},\ \varphi_{e,w} = \frac{\mathrm{N_{e,w}} + \iota_w}{\sum_{\mathrm{w}}^{\mathrm{W}}(\mathrm{N_{e,w}} + \iota_w)}$$
$$\mu_{d,x} = \frac{K_{d,x} + \epsilon_x}{\sum_{\mathrm{x}}^{\mathrm{X}}(\mathrm{K_{d,x}} + \epsilon_x)},\ \Omega_{e,u} = \frac{\mathrm{C_{e,u}} + \eta_u}{\sum_{\mathrm{u}}^{\mathrm{U}}(\mathrm{C_{e,u}} + \eta_u)}$$

4. EXPERIMENT

In the previous section we introduced the TUT model, a new statistical model for topic and trend tracking and user interest discovery. In this section, we evaluate the effectiveness of our

model with experiments on three different datasets. First of all, we conducted a qualitative analysis on the topics and trends detected by our model. Secondly, we compared our model to the Trend Analysis Model [9] on several different datasets by using perplexity and rate of convergence as measurements. Finally, we built a recommendation model to test if the trend-user distribution is effective to predict user participation.

4.1 Data sets

Data 1 (**DBLP Dataset**): We implemented a XML parser to extract bibliography data from DBLP. Our dataset contains 10 years (2001-2010) of research papers in conference proceedings including CIKM, KDD, WWW and SIGIR. Each document in this dataset corresponds to a publication where title, author list and year of publication were extracted to form the content, user list and timestamp respectively. We preprocessed the data by removing stop words, stemming, and filtering low frequency words and authors. Concretely, we obtained a total set of 4,631 documents and 1,627 users. We considered one year as the time interval. The training set includes documents from 2001 to 2009 and the test set consists of documents from 2010.

Data 2 (**Digg Dataset**): Digg (digg.com) is a social news website where users can post news, make comments, or vote up/down other users' posts/comments. In this experiment, we selected the 5 most popular news sources in Digg: CNN, BBC, NPR, The Washington Post, and Yahoo! News. We then built our dataset by using Digg open API to collect all posts as well as their comments from these five news sources during the period between March 1st 2011 and May 31st 2011. Each post and all of its comments were aggregated as a document in this dataset. Contents from both a post and its comments were combined to be the content of a document. The post initiator and all commenters were extracted to be a user list. The same pre-processing used for DBLP dataset was then applied. Accordingly, we obtained a total set of 9,894 documents and 2,356 users. We considered a week as the time interval. The training set includes the documents from March 1st 2011 to May 15th 2011 and the test set consists of the documents from May 16th 2011 to May 31st 2011.

Data 3 (**MySpace Dataset**): MySpace is one of the popular social networking sites which offers its registered users to start a new thread to discuss several topics or participate in a thread created by other users and make comments. In this study, we used a public data set available for workshop CAW 2.0 (http://caw2.barcelonamedia.org/). The original MySpace dataset provided by the workshop consists of threads from three different sub-forums: campus life, news & politics and movies. In our experiment, we used threads from news & politics sub-forum and movies sub-forum. The same pre-processing used for DBLP dataset was then applied. We transformed the original dataset to an input format of our model. Each thread and its comments were aggregated as a document in our dataset. Contents from both an initial post and its comments were combined as the content of a document. The post initiator and all commenters were extracted to be the user list. Accordingly, we obtained a total set of 3,886 documents and 1,373 users. The time range of our processed dataset spans from October 2007 to November 2008. Therefore, we considered a month as the time interval. The training set includes the documents from October 2007 to September 2008 and the test set consists of the documents from October 2008 to November 2008.

4.2 Case Study

In the first experiment, we conducted a qualitative analysis on the **MySpace dataset** to investigate the quality of the topics and trends detected by our proposed model. In order to produce a small example fitting this paper's size, we chose the number of topic $|Z|= 50$. We empirically set $\alpha = 1$ and let $\gamma = 0.1, \eta = 0.1, \beta = 0.5, \iota = 0.1, \epsilon_0 = 0.1, \epsilon_1 = 0.2$ and $\epsilon_3 = 0.7$. Table 2 shows an illustrative example of 10 different topics extracted from MySpace dataset of August 2008 by our proposed model. The top 10 words with the largest topic-word association were presented for each topic in table 2. The top 10 words in each topic in table 2 show clearly what each topic is about and how the topics are differentiated. For instance, topic 6 is about taxation and

Table 2. Discovered Topics from MySpace Dataset

Topic 0	Topic 3	Topic 6	Topic 11	Topic 14
Russia 0.0605	Georgia 0.1015	money 0.0142	US 0.0224	vote 0.0198
war 0.0157	Russian 0.0578	pay 0.013	Iraq 0.0224	election 0.0137
US 0.0139	Ossetia 0.0271	people 0.0107	war 0.0123	power 0.0129
country 0.0102	south 0.0242	job 0.0107	al 0.0114	party 0.0128
world 0.0098	Saakashvili 0.0130	wage 0.0100	force 0.0111	state 0.0122
American 0.0083	peace 0.0075	company 0.0098	military 0.0106	president 0.0098
Georgia 0.0078	army 0.0072	work 0.0082	Pakistan 0.0098	republican 0.0094
soviet 0.0072	attack 0.0070	market 0.0080	terrorist 0.0091	country 0.0086
state 0.0070	force 0.0066	tax 0.0080	fight 0.0090	govern 0.0084
military 0.0062	troop 0.0065	price 0.0078	Afghanistan 0.0088	democrat 0.0075
Topic 26	Topic 32	Topic 39	Topic 40	Topic 49
kill 0.0278	American 0.0277	finance 0.0203	Obama 0.0467	bank 0.0388
Afghanistan 0.0207	Bush 0.0141	US 0.0155	McCain 0.0380	bailout 0.0232
civilian 0.0203	world 0.0094	mortgage 0.0123	Palin 0.0171	hometown 0.0144
war 0.0142	US 0.0083	loan 0.0095	vote 0.0133	loan 0.0111
bomb 0.0108	debt 0.0068	fund 0.0092	campaign 0.0084	bail 0.0078
Taliban 0.0091	George 0.0055	asset 0.0087	Bush 0.0084	mortgage 0.0075
UN 0.0083	European 0.0052	fed 0.0085	republican 0.0072	north 0.0061
attack 0.0080	critic 0.0053	institute 0.0082	president 0.0069	street 0.0061
British 0.0080	leader 0.0045	govern 0.0076	democrat 0.0067	crisis 0.0056
news 0.0078	student 0.0045	treasury 0.0076	Barack 0.0064	Bush 0.0052

employment. Topic 11 is about military. Topic 39 is about finance and topic 14 is about campaign.

We further analyzed the detected trends by selecting three trends as examples and listing top 5 topics with the largest trend-topic association for each trend, as showed in table 3. These three trends were manually labeled by human annotator according to the content of the threads assigning to them and their top topic list. These three detected trends are: 2008 Georgia-Russia Crisis, 2008 Mortgage and Credit Crisis and 2008 U.S. Presidential Election. As shown in table 3, some topics were shared by these three trends. We highlighted a topic in blue if it was shared by two trends and highlighted a topic in red if it was shared by three trends. For example, from table 3, we noticed that topic 14 (campaign topic), was shared by all three trends. From our understanding, trend A (2008 Georgia-Russia Crisis) happened around the same time as trend C (2008 U.S. Presidential Election). At that moment, presidential candidates' attitude toward this crisis was critical to exhibit their experience and capability of dealing with foreign affair, which explains why threads of trend A involved in campaign topic. Similarly, 2008 Mortgage and Credit Crisis continued to worsen from March 2008, when the U.S. Government freed up $200 billion to support Fannie Mae and Freddie Mac mortgage giants, to September 2008, when Lehman Brothers investment bank declared bankruptcy and the U.S. Government bailed out AIG. As a result, threads of trend B (2008 Mortgage and Credit Crisis) unavoidably involved campaign topic which showed the American's hope for a new president to remedy the economic problems. Similarly, threads of trend C (2008 U.S. Presidential Election) shared topic 6 (taxation and employment) and topic 39 (finance) with trend B (2008 Mortgage and Credit Crisis), which is also reasonable because possible solutions to the Mortgage and Credit Crisis were very hot topic during the 2008 U.S. Presidential Election. To sum up, without modeling trends and their associated topics (as in the traditional topic modeling works), it is difficult for us to a gain comprehensive view of each trend and find out the connections among them through their common topics.

In addition, we measured the volume of threads in each trend in one year period and visualized their accumulated volume in Figure 4. X-axis in figure 4 represents timeline and y-axis of figure 4 indicates the accumulated percentage of volume of the whole dataset. Concretely, 2008 U.S. Presidential Election was the most eye-catching event which attracted lots of threads in MySpace between Oct 2007 and Oct 2008. This campaign involved several different issues including taxation, employment, finance, counter-terrorism, military and more. We observed two peaks of trend C (green area). The first peak happened around February 2008 which can be explained by the primaries and caucuses, especially on the Super Tuesday, 5 February 2008. The second peak happened when the Election Day was getting closer. Similarly, for the trend of 2008 Mortgage and Credit Crisis, its volume was close to zero before 2008 but became substantially larger since January 2008. As we know, the Mortgage and Credit Crisis continued to worsen since UBS reported an $18 billion write-down due to its exposure to the American real estate market in January 2008 and Fannie Mac reported $3.55 billion loss for the fourth quarter of 2007 in February 2008. The increase of trend B's volume also can be explained by the fact that the crisis became even worse and the U.S. Congress gave final message to multi-billion-dollar program and tried to address the crisis in July 2008. Further, trend A kept going upward which indicated the Georgia-Russia crisis was getting more and more intense, especially when the South Ossetia War broke out on 7 August, 2008.

Table 3. Top Topics of Each Discovered Trend

Trend	Popular Topic List
A. 2008 Georgia-Russia Crisis	0,3,26,14,40
B. 2008 Mortgage and Credit Crisis	6,32,14,39,49
C. 2008 U.S. Presidential Election	40,14,6,39,11

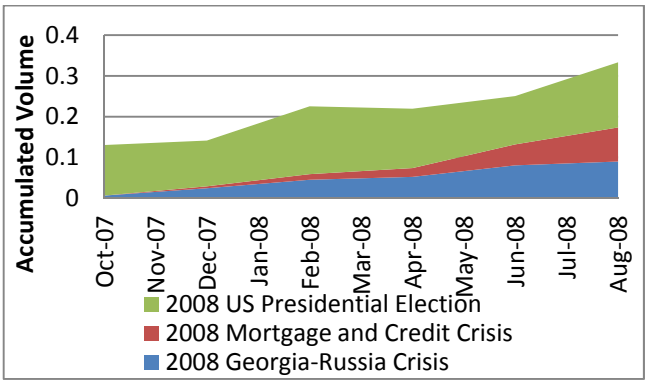


Figure 4. Accumulated Volume of Trends in MySpace Dataset

4.3 Quantitative Evaluation

In this section, we used perplexity as an evaluation metric to investigate the performance of our proposed model in 3 different datasets. For each dataset, we trained the model in its training set and computed its perplexity in the test dataset. Trend Analysis Model (TAM) [9] is employed as a benchmark. The major differences between TUT and TAM are: 1) TUT can automatically decide the number of trends while TAM needs human intervention to predefine the number of trends; 2) TAM is unsuitable for processing document streaming; and 3) TUT models the generation processes of contents and user participations in a unified framework while TAM neglects all user information.

4.3.1 Evaluation Criterion

Perplexity is a standard measure for evaluating the generalization performance of a probabilistic model. The value of perplexity reflects the ability of a model to generalize to unseen data. Specifically, in our case, perplexity reflects the ability of a model to predict words for unseen documents. The perplexity is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates better generalization performance. We follow [9] and define the perplexity score for a test set D_{test} as:

$$PPX(D_{test}) = \exp\left(-\frac{1}{W}\sum_{d=1}^{|D_{test}|}\sum_{w\in d}^{|d|}\log\left(\Pr\left(w_d\right)\right)\right)$$

where

$$\Pr(w_d) = \mu_{d,0}\phi_{bg,w} + \mu_{d,1}\varphi_{e_d,w} + \sum_{z}^{Z}\mu_{d,2}\theta_{e_d,z}\phi_{z,w}$$

In the equation above, $\mu_{d,i}$ represents the probability of observing switch variable equals to i in document d, $\phi_{z,w}$ represents the probability of observing word w given topic z, $\varphi_{e_d,w}$ represents

the probability of observing word w given trend e_d, and $\theta_{e_d,z}$ represents the probability of observing topic z given trend e_d. The probabilities $\phi_{bg,w}$, $\phi_{z,w}$, $\varphi_{e_d,w}$ and $\theta_{e_d,z}$ (when e_d is an existing trend in the training set) are learned from the training set, and $\mu_{d,0}$, $\mu_{d,1}$, $\mu_{d,2}$, $\varphi_{e_d,w}$ and $\theta_{e_d,z}$ (when e_d is a new trend in the test set) are estimated by a Gibbs sampling process on the test set.

4.3.2 *Effect of Recurrent Chinese Restaurant Process*

As mentioned in the literature review, TAM needs human intervention to predefine its number of trends |C|. However, in real world application, it's nontrivial to find out the optimal |C|. In addition, our proposed model can automatically decide the number of trends in a dataset without human intervention. In this sub-section, we first studied the performance of TUT on **Digg dataset**. For TUT model, we set $\alpha = 1$, $\gamma = 0.1, \eta = 0.1, \beta = 50/|Z|, \iota = 0.1, \epsilon_0 = 0.1, \epsilon_1 = 0.2$ and $\epsilon_3 = 0.7$. The experiment results were shown in figure 5. It's important to note that both TAM and TUT have fixed number of topics which needs to be predefined. Therefore, the x-axis in Figure 5 represents different numbers of topic, |Z|, from 50 to 300. Y-axis in Figure 5 represents the perplexity score in test set. TAM(5) denotes a TAM model with predefined number of trend |C|=5. TAM(10) denotes a TAM with |C|=10 and TAM(20) denotes a TAM with |C|=20. The number of trends determined automatically by TUT is around 11 (average number of 5 repeating experiments), so that we only tested TAM with |C| equaling to 5, 10 and 20. Figure 5 shows that TUT model achieved the lowest perplexity score across all settings of topic number comparing to TAM. We also observed that in TAM the perplexity score roughly increased with the value of |C| while TUT performed relatively stable. Last but not least, even though TAM(10) has a very close trend number to TUT in Digg dataset, TUT still outperformed TAM(10).

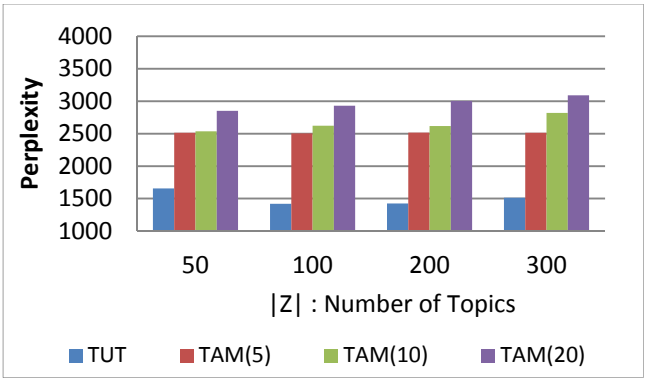


Figure 5. Perplexities over different # of topics for TUT and TAM model with different # of trend in Data 2

Furthermore, we compared the change of perplexity over iteration number during Gibbs sampling process. The results are shown in Figure 6. We set the number of topics, |Z|, to be 100 based on the following considerations: 1) all four models achieved decent performances when |Z| = 100; and 2) when |Z| = 100, it simplifies the result for clear visualization. From this figure, we observed that all these models are able to converge quickly. But given a similar starting point, there is a larger decrease for TUT's perplexity score comparing to TAM(5), TAM(10) and TAM(20)

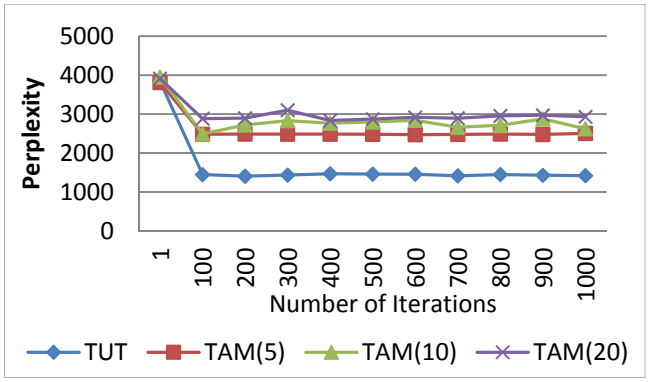


Figure 6. Perplexities over the number of iterations for TUT model and TAM model with different # of trend in Data 2

4.3.3 *Performance of TUT in all other Datasets*

Additionally, we compared the performance of TUT and TAM in the other two datasets. We repeated each model in each setting for five times and reported the average number. The experiment results were summarized in Figure 7 and Figure 8 respectively. For DBLP dataset, we predefined the number of trends for TAM to be 5, 10 and 20 respectively, because the average number of trends identified by our proposed model in test set equals to 11.5. We set $\alpha = 1$, $\gamma = 0.1, \eta = 0.1, \beta = 50/|Z|, \iota = 0.1, \epsilon_0 = 0.1, \epsilon_1 = 0.2$ and $\epsilon_3 = 0.7$. Similarly, we predefined the number of trends for TAM in MySpace dataset to be 5, 10 and 20 respectively since the average number of trends identified by our proposed model in MySpace's test set equals to 20.25. We set $\alpha = 0.1, \gamma = 0.1, \eta = 0.1, \beta = 50/|Z|, \iota = 0.1, \epsilon_0 = 0.1, \epsilon_1 = 0.2$ and $\epsilon_3 = 0.7$.

From both Fig. 7 and Fig. 8, we observed that our proposed TUT model consistently outperformed TAM in all settings of trend number and topic number. In addition, the perplexity scores of TUT were relatively stable comparing to TAM. As we know, the number of trends should be different in different time interval, but we predefined and fixed the number of trends for TAM according to the result of TUT, which might bring difficulties for TAM to generalize to the test dataset. However, even though TAM(10) has a very close number of trends to TUT in DBLP dataset, TUT still outperformed TAM(10). The same applies to TAM(20) in MySpace dataset. One possible explanation is that, modeling user participation, to some extent, might help TUT model to better generalize to the unseen dataset.

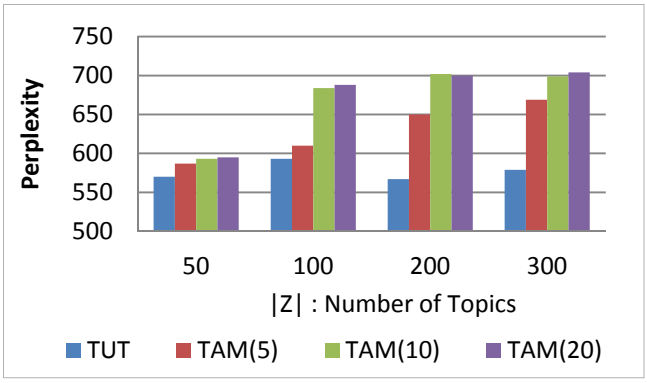


Figure 7. Perplexities over different number of topics for TUT and TAM model with different # of trend in Data 1

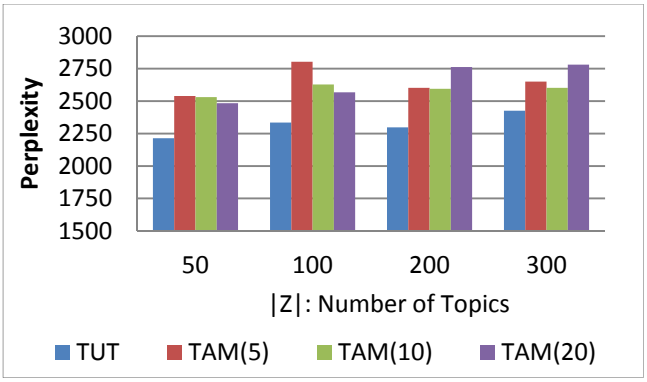


Figure 8. Perplexities over different number of topics for TUT and TAM model with different # of trend in Data 3

4.4 User Prediction

Modeling user interest is critical for news/blog recommendation and personalized search. One of the important outputs of TUT model is trend-user distributions, which can be considered as users' interests toward each detected trend. In this experiment, we evaluated the usefulness of the user interests to boost ranking for potential participant prediction. The experiment was conducted in **Digg dataset**. Generally speaking, we first trained a recommendation model based on the Digg dataset' training set. Then, for each document in its test set, we randomly removed 80% of the users and only withheld the other 20%. We then used the recommendation model to predict users and compared the result to the ground truth (the removed users). We first used a naïve ranking method as a baseline approach and then boosted the baseline by the user interests detected by our proposed model. Our goal is to show that, even though with a straightforward combination method, the detected user interests can boost the naïve ranking approach substantially.

Specifically, for the **naïve ranking approach**, we first constructed a bipartite graph based on the user-document relationships observed in the training dataset. The bipartite graph consists of nodes and edges. Each node represents either a Digg user or a document posted in Digg.com. All nodes representing users are on the left hand side while all nodes representing documents are on the right hand side. Each edge of the bipartite graph connects a user to a document which means that a user contributed to a document. Given this bipartite graph, the similarity between user i and j is defined by the Jaccard Index as: $Sim(i,j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}$, where C_i represents the set of documents that user i participated. In this way, we calculated the similarity between every two users in the training set. Given a new document d in the test set and its 20% of withheld users, $\overline{P_d}$, we ranked all other users, excluding those in $\overline{P_d}$, based on $Pr(\overline{\overline{P_{d_i}}}|\overline{P_d}, \mathcal{G})$ defined as:

$$Pr(\overline{\overline{P_{d_i}}}|\overline{P_d}, \mathcal{G}) = \frac{\sum_{j \in \overline{P_d}} |Sim(i,j)|}{|\overline{P_d}|}$$

where $\overline{\overline{P_{d_i}}}$ represents a user i who does NOT exist in $\overline{P_d}$, and $\mathcal{G}$ represents a bipartite graph. The rationale of this naïve raking is that, if a user is very similar to those existing users of a given document, then this user is also likely to exist in this document. This naïve ranking method relying on user similarity will serve as the baseline approach in this experiment.

On the other hand, for our **boosted ranking method**, based on the trend-user distributions Ω that we obtained from the training set, we know which trends a user will be interested. By applying a trained TUT model to the test set, it automatically assign each document in the test set to a trend e. Since we already know users' interests toward each trend e, $\Omega_e = \{\Omega_{e,1}, \Omega_{e,2}, \ldots, \Omega_{e,u}\}$ from the training set, the boosted ranking for user $\overline{\overline{P_{d_i}}}$ to document d, is then redefined as:

$$\Pr(\overline{\overline{P_{d_i}}}|\overline{P_d}, \mathcal{G}, e_d, \Omega_{e_d}) = w \times \Pr(\overline{\overline{P_{d_i}}}|e_d, \Omega_{e_d}) + (1-w) \times \Pr(\overline{\overline{P_{d_i}}}|\overline{P_d}, \mathcal{G})$$

where $\Pr(\overline{\overline{P_{di}}}|e_d, \Omega_{e_d})$ equals to $\Omega_{e_d, \overline{\overline{P_{d_i}}}}$, representing user i's interest to trend e_d and w denotes the weight to combine the baseline approach and user interests.

The experiment setup is as follows: 1) Using the training set, we constructed a user-thread bipartite graph and calculated user similarities; 2) Based on the user similarities computed in step 1, we applied the baseline ranking method to predict users for each document in the test set (results from the first two steps were used as baseline); 3) Using the training set again, we trained our TUT model to discover users' interests to trends; 4) We applied the trained TUT model to assign a trend label to each document in the test set; 5) For each document in the test set, since we knew its trend label from step 4 and also users' interests to each trend from step 3, we applied our boosted ranking method to predict its potential participants.

4.4.1 Measurement

We employed Precision@K as the evaluation measure in this experiment. Specifically, by using Precision@K, we consider the top-K list resulted by a recommendation method. Let $\boldsymbol{m}$ denotes the number of users that co-exist in the top-K list and the ground truth list of a document, Precision@k equals to $\boldsymbol{m/K}$. The higher the value, the better the performance of a recommendation method is.

4.4.2 Result

We empirically set weight w to 0.5. As shown in Table 4, our boosted ranking method consistently outperformed the naïve ranking from K=5 to K=30. The highlighted red figures within parentheses indicate the percentage of improvement of boosted rank comparing to the baseline approach.

Table 4. Precision@K of Baseline and Boosted Ranking

	P@5	P@10	P@15	P@20	P@25	P@30
Naïve Ranking	0.069	0.067	0.062	0.060	0.058	0.055
Boosted Ranking	0.081 (17%)	0.075 (12%)	0.074 (19%)	0.070 (17%)	0.068 (17%)	0.062 (12%)

5. CONCLUSION

In this paper, we proposed a Topic-User-Trend Model to simulate the generation process of user-generated web contents. By incorporating a latent variable "trend" and using RCRP to enable our model to decide automatically the appropriate number of trends in a document corpus and also taking users' interests into account, our proposed model is able to model the generation process of web contents in a more meaningful way. At the same time, our model achieves a better generalization performance than the TAM model in multiple test datasets. Besides, the user interests learned by our proposed model can be utilized for document recommendation, which substantially improved the performance of the naïve ranking method which relies on user

similarity only. Thus far, our proposed model only considers users participated in a given document without taking their interactions into consideration. In the future work, we will incorporate the generation process of user relationships into our model. Moreover, we will further explore other opportunities to make use of the trend-user association for recommendation problems.

6. REFERENCES

[1] Ahmed, A., Ho, Q., Eisenstein, J., Xing, E., Smola, A. J. and Teo, C. H. 2011. Unified analysis of streaming news. Proceedings of the 20th international conference on World Wide Web (WWW'11) ACM 267-276.

[2] Ahmed, A. and Xing, E. 2008. Dynamic non-parametric mixture models and the recurrent chinese restaurant process. Proceedings of SDM 2008.

[3] 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. Proceedings of the 2008 Eighth IEEE International Conference on Data Mining IEEE Computer Society 3-12.

[4] Blei, D., Ng, A. and Jordan, M. 2003. Latent dirichlet allocation. The Journal of Machine Learning Research. 3, 993-1022.

[5] Blei, D. M. and Lafferty, J. D. 2006. Dynamic topic models. Proceedings of the 23rd international conference on Machine learning Pittsburgh, Pennsylvania ACM 113-120. http://doi.acm.org/10.1145/1143844.1143859

[6] He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P. and Giles, L. 2009. Detecting topic evolution in scientific literature: how can citations help? Proceeding of the 18th ACM conference on Information and knowledge management ACM 957-966.

[7] Hearst, M. A. and Pedersen, J. O. 1996. Reexamining the cluster hypothesis: scatter/gather on retrieval results. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval Zurich, Switzerland ACM 76-84. 10.1145/243199.243216

[8] Hofmann, T. 1999. Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval ACM New York, NY, USA 50-57.

[9] Kawamae, N. 2011. Trend analysis model: trend consists of temporal words, topics, and timestamps. Proceedings of the fourth ACM international conference on Web search and data mining ACM 317-326.

[10] McCallum, A., Corrada-Emmanuel, A. and Wang, X. 2005. Topic and role discovery in social networks. Proceedings of the 19th international joint conference on Artificial intelligence Morgan Kaufmann Publishers Inc. 786-791.

[11] Mei, Q. and Zhai, C. 2005. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining ACM New York, NY, USA 198-207.

[12] Morinaga, S. and Yamanishi, K. 2004. Tracking dynamics of topic trends using a finite mixture model. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining Seattle, WA, USA ACM 811-816. http://doi.acm.org/10.1145/1014052.1016919

[13] Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P. 2004. The author-topic model for authors and documents. Proceedings of the 20th conference on Uncertainty in artificial intelligence Banff, Canada AUAI Press 487-494.

[14] Schult, R. and Spiliopoulou, M. 2006. Discovering emerging topics in unlabelled text collections. Lecture Notes in Computer Science. 4152, 353-366.

[15] Shahaf, D., Guestrin, C. and Horvitz, E. 2012. Trains of thought: generating information maps. Proceedings of the 21st international conference on World Wide Web Lyon, France ACM 899-908.

[16] Sun, Y., Tang, J., Han, J., Gupta, M. and Zhao, B. 2010. Community evolution detection in dynamic heterogeneous information networks. Proceedings of the Eighth Workshop on Mining and Learning with Graphs ACM 137-146.

[17] Wang, X. and McCallum, A. 2006. Topics over time: a non-markov continuous-time model of topical trends. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining ACM 424-433.

[18] Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y. and Ma, J. 2004. Learning to cluster web search results. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval Sheffield, United Kingdom ACM 210-217. 10.1145/1008992.1009030