A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs

Qiaozhu Mei[†], Chao Liu[†], Hang Su[‡]
ChengXiang Zhai[†]
[†]Department of Computer Science
University of Illinois at Urbana-Champaign
[‡]Department of EECS
Vanderbilt University

ABSTRACT

Mining subtopics from weblogs and analyzing their spatiotemporal patterns have applications in multiple domains. In this paper, we define the novel problem of mining spatiotemporal theme patterns from weblogs and propose a novel probabilistic approach to model the subtopic themes and spatiotemporal theme patterns simultaneously. The proposed model discovers spatiotemporal theme patterns by (1) extracting common themes from weblogs; (2) generating theme life cycles for each given location; and (3) generating theme snapshots for each given time period. Evolution of patterns can be discovered by comparative analysis of theme life cycles and theme snapshots. Experiments on three different data sets show that the proposed approach can discover interesting spatiotemporal theme patterns effectively. The proposed probabilistic model is general and can be used for spatiotemporal text mining on any domain with time and location information.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Text Mining

General Terms: Algorithms

Keywords: Spatiotemporal text mining, weblog, mixture model, theme pattern

1. INTRODUCTION

With the quick growth during recent years, *weblogs* (or *blogs* for short) have become a prevailing type of media on the Internet [7]. Simultaneously, increasingly more research work is conducted on weblogs, which considers blogs not only as a new information source, but also as an appropriate testbed for many novel research problems and algorithms [16, 26, 11, 10]. We consider weblogs as online diaries published and maintained by individual users, ordered chronologically with time stamps, and usually associated with a profile of their authors. Compared with traditional media such as online news sources (e.g., CNN online) and public websites maintained by companies or organizations (e.g., Yahoo!), weblogs have several unique characteristics: 1) The content of weblogs is highly personal and rapidly evolving. 2) Weblogs are usually associated with the personal information of their authors, such as age, geographic location and personal interests [17]. 3) The interlinking structure of weblogs usually forms localized micro communities, reflecting relations such as friendship and location proximity [17].

With these characteristics, weblogs are believed to be appealing for research across multiple domains to answer questions such as "what happens over time", "how communities are structured and evolving", and "how information diffuses over the structure". Specifically, there are currently two major lines of research on blog analysis. One line is to understand the interlinking structures (i.e. communities) and model the evolution of these structures. Kumar and others [17] introduced the distribution of blogs over locations and studied how they form communities. They also proposed a way to discover bursty evolution of these communities [16]. The other line is to perform temporal analysis on blog contents and model information diffusions among blogs. Gruhl and others [11] proposed a model for information propagation and categorized diffusing topics into chatter and spikes; they followed up to prove these temporal patterns of topics are useful to predict spike patterns in sales ranks [10].

Although these existing studies have addressed some special characteristics of weblogs, such as community structures, rapid evolution and time stamps, none of them has addressed well the following two needs in analyzing weblogs.

1. Modeling mixture of subtopics: The content of weblogs often includes personal experiences, thoughts and concerns. As a result, a blog document often contains a mixture of distinct subtopics or themes. For example, a blog article about the topic "Hurricane Katrina" may contain different aspects of concerns (i.e., different themes) such as "oil price" and "response of the government" or even some other topics. In many applications, extracting and analyzing such themes of an event are highly desirable. For example, a news analyzer may ask "what are the growing concerns of common people about Hurricane Katrina", while a marketing investigator may be interested in "what features of iPod do people like or dislike most". To answer such questions, we must analyze the *internal* theme components within a blog article. Unfortunately, most existing work (e.g., [11, 8]) assigns a blog article to only one topic, which has not attempted to model such a mixture of subtopics within a blog document and is thus unable to perform accurate fine-granularity subtopic analysis.

2. Spatiotemporal content analysis: In addition to the available time stamps, a considerable proportion of weblogs are also associated with the profiles of their authors

Copyright is held by the International World Wide Web Conference Committee (IW3C2). Distribution of these papers is limited to classroom use, and personal use by others.
WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

which provide information about their geographic locations. In general, the content information of weblogs may be related with or depend on both the time stamp of the article and the location of its author. For example, when analyzing the spikes of topics, it is very likely to have a considerable gap of time between the spikes of discussion on the new book "Harry Potter" in England and in China. Therefore, using the average temporal pattern of the book discussion to predict the sales in a specific location would not be reasonable. The public opinions may also be location-dependent. For example, people outside the United States may be concerned more about "shipping price" and "repair warranty" of IBM laptops than those inside the United States, and the public responses and concerns of Hurricane Katrina may appear differently between Louisiana and Illinois.

Due to the inherent interaction of the content of weblogs with both time and location, it is highly desirable to analyze weblogs in a temporal and spatial context. Indeed, many interesting questions could only be answered by connecting content with time and location and analyzing spatiotemporal theme patterns. For example, a sociologist may ask "what do people in Florida think about the presidential candidates and how do their opinions evolve over time", while a business provider may be interested in comparing the customer responses to their new product from two countries.

Although some previous work (e.g., [17]) has considered temporal and spatial information associated with weblogs, no previous work has addressed well the need for correlating the content, especially the multiple themes *within* articles, with spatiotemporal information.

In this paper, we study the novel problem of discovering and summarizing the spatiotemporal theme patterns in weblogs. We formally define this problem and propose a novel probabilistic method for modeling the most salient themes from a text collection and their distributions and evolution patterns over time and space. We evaluate the proposed mixture model on three different data sets – collections of blog entries from MSN Space about "Hurricane Katrina", "Hurricane Rita", and "iPod Nano".

The results show that our method can effectively discover major interesting themes from text and model their spatiotemporal distributions and evolutions. The mining results can be used to further support higher level analysis tasks such as user behavior prediction, information diffusion and blogspace evolution analysis.

The proposed method is completely unsupervised and can be applied to any text collection with time stamps and location labels, such as news articles and customer reviews. The method thus has many potential applications, such as **(1) Search result summarization:** Provide a summary for blogsearch results, which consists of themes, snapshots of spatial distributions of themes, and temporal evolution patterns of themes. **(2) Public opinion monitoring:** Extract the major public concerns for a given event, compare the spatial distributions of these concerns, and monitor their changes over time. **(3) Web analysis:** Extract major themes and model the macro-level information spreading and evolution patterns on the blogspace. **(4) Business intelligence:** Facilitate the discovery of customer opinions/concerns and the analysis of their spatial distributions and temporal evolutions.

The rest of the paper is organized as follows. In Section 2, we formally define the general problem of spatiotemporal theme pattern discovery. In Section 3, we present probabilistic mixture models to model themes over time and space. We discuss our experiments and results in Section 4. Finally, we discuss related work in Section 5 and conclude in Section 6.

2. PROBLEM FORMULATION

The general problem of spatiotemporal theme pattern discovery can be formulated as follows.

Formally, we are given a collection of text documents with time stamps and location labels, $C = \{(d_1, \tilde{t}_1, \tilde{l}_1), (d_2, \tilde{t}_2, \tilde{l}_2), ..., (d_n, \tilde{t}_n, \tilde{l}_n)\}$, where $\tilde{t}_i \in T = \{t_1, t_2, ..., t_{|T|}\}$ and $\tilde{l}_i \in L = \{l_1, l_2, ..., l_{|L|}\}$ are the time stamp and location label of document d_i respectively. Each document is a sequence of words from a vocabulary set $V = \{w_1, ..., w_{|V|}\}$.

In information retrieval and text mining, it is quite common to use a word distribution to model topics, subtopics, or themes in text[3, 12, 1, 21]. Following [21], we define a theme as follows:

Definition 1 (Theme) A ***theme*** in a text collection C is a probabilistic distribution of words characterizing a semantically coherent topic or subtopic. Formally, a theme is represented with a (theme-specific) unigram language model θ, i.e., a word distribution $\{p(w|\theta)\}_{w \in V}$ s.t. $\sum_{w \in V} p(w|\theta) = 1$.

High probability words of such a distribution often suggest what the theme is about. For example, a theme about "oil price" associated with the topic Hurricane Katrina may have high probabilities for words like "oil", "price", "gasoline", "increase", etc. In this definition, we assume that the basic meaningful unit of text is a word, which is generated independently of each other. This definition can be generalized to adopt multi-word phrases as meaningful units.

To model the temporal patterns of themes, we define the concept "theme life cycle" as follows:

Definition 2 (Theme Life Cycle) Given a theme represented as language model θ, a location l and a set of consecutive time stamps $T = \{t_1, t_2, ..., t_{|T|}\}$, the ***Theme Life Cycle*** of theme θ at location l is the conditional probability distribution $\{P(t|\theta, l)\}_{t \in T}$. Clearly, $\sum_{t \in T} P(t|\theta, l) = 1$. We define the ***Overall Life Cycle*** of a theme as $\{P(t|\theta)\}_{t \in T}$ if no specific location is given.

A theme life cycle can be visualized by plotting the density function of the time-theme distribution $\{P(t|\theta, l)\}_{t \in T}$ continuously over the entire T.

In our previous work [21], the life cycle of a theme is defined as the theme's strength spectrum over the whole time line [21], which does not have a probabilistic interpretation. Here we give a probabilistic definition of the life cycle so that it can be involved in probabilistic models. Note that under our definition, the strength of different themes is not directly comparable. That is, given a time $\tilde{t}$, $P(\tilde{t}|\theta_1, l) > P(\tilde{t}|\theta_2, l)$ does not necessarily imply that θ_1 is stronger than θ_2 at $\tilde{t}$. To compare the strength of θ_1 and θ_2, we can compute $P(\theta_1|\tilde{t}, l)$ and $P(\theta_2|\tilde{t}, l)$ using Bayes rule.

To model spatial patterns, we further define the "theme snapshot" of the collection at a given time.

Definition 3 (Theme Snapshot) Given a set of themes represented as language models $\Theta = \{\theta_1, \theta_2, ...\theta_k\}$, a time stamp t and a set of locations $L = \{l_1, l_2, ..., l_{|L|}\}$, the ***Theme Snapshot*** at time t is defined as the joint probability distribution of θ and l conditioned on t, i.e. $\{P(\theta, l|t)\}_{\theta \in \Theta, l \in L}$. Naturally, we have $\sum_{\theta \in \Theta, l \in L} P(\theta, l|t) = 1$.

A theme snapshot can be visualized by a map of theme distributions over all locations for a given time. Note that

with this definition, the strength of two themes at the same location is directly comparable, i.e. given a location $\tilde{l}$, if $P(\theta_1, \tilde{l}|t) > P(\theta_2, \tilde{l}|t)$, we have that θ_1 is stronger than θ_2 at location $\tilde{l}$ during the time period t.

Intuitively, we may assume that some global themes would span the whole collection. Given a text collection C with time and location labels, we define the major tasks of the **Spatiotemporal Theme Pattern (STTP)** discovery problem as follows: 1) automatically extract a set of major themes from C; 2) for a given location, compute the life cycles of the common themes at this location; 3) for a given time period, compute the theme snapshot over all locations.

Formally, given $C = \{(d_1, \tilde{t}_1, \tilde{l}_1), (d_2, \tilde{t}_2, \tilde{l}_2), ..., (d_n, \tilde{t}_n, \tilde{l}_n)\}$, the task of STTP discovery is to: 1) discover global themes $\Theta = \{\theta_1, ..., \theta_k\}$; 2) for each given global theme θ and location l, compute $P(t|\theta, l)$ for all $t \in T$; 3) for each given t, compute $P(\theta, l|t)$ for all $\theta \in \Theta$ and $l \in L$.

STTP discovery is challenging for several reasons. First, in general, no training data is available to discriminate themes, thus we have to rely on completely unsupervised methods to discover STTP. Second, themes are latent in the collection and usually associated with the same event or topic. Therefore, existing techniques of novelty detection and event tracking, which aim to segment the text and find the boundaries of events [2, 25, 19], cannot be applied directly to this problem. Finally, a unified analysis of theme life cycles and theme snapshots requires a careful design of models so that we can explain the generation and evolution of themes over time and space in a completely unsupervised way.

In the next section, we present a unified probabilistic approach for discovering STTPs.

3. PROBABILISTIC SPATIOTEMPORAL THEME ANALYSIS

3.1 General Idea

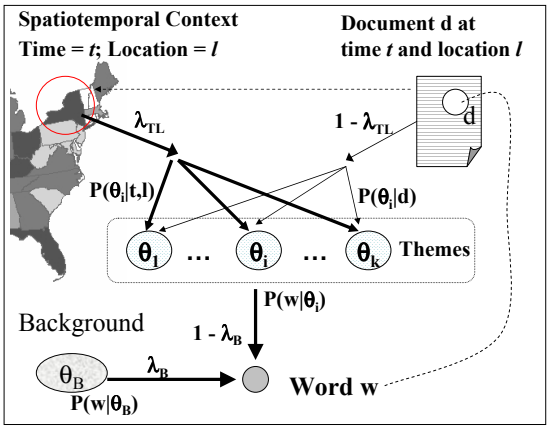


Figure 1: The generation process of a word in the spatiotemporal theme model

Previous work has shown that mixture models of multinomial distributions (i.e., mixture language models) are quite effective in extracting themes from text [12, 1, 9, 27, 21]. The basic idea of such approaches is to assume that each word in the collection is a sample from a mixture model with multiple multinomial distributions as components, each representing a theme. By fitting such a model to text data, we can obtain the distributions for the assumed themes.

Our main idea for probabilistic spatiotemporal theme analysis is to adopt a similar approach to extract themes and extend existing work on mixture models to incorporate a time variable and a location variable. Intuitively, the words in a blog article can be classified into two categories: (1) common English words (e.g., "the", "a", "of"); (2) words related to the global subtopics (themes) whose spatiotemporal distribution we are interested in analyzing (e.g., "Hurricane Katrina and oil price"). Correspondingly, we introduce two kinds of theme models: (1) θ_B is a background theme model to capture common English words; (2) $\Theta = \{\theta_1, ..., \theta_k\}$ are k global themes to be used for all articles in the collection.

With these theme models, a document of time t and location l can be modeled as a sample of words drawn from a mixture of k global themes θ_1, ..., θ_k and a background theme θ_B. To model spatiotemporal characteristics of themes, we assume that the theme coverage in a document depends on the time and location of the document. The mixture model is illustrated in Figure 1.

Such a mixture model can be interpreted as modeling the following process of "writing" a weblog article: An author at time t and location l would write each word in the article by making the following decisions stochastically: (1) The author would first decide whether the word will be a non-informative English word, and if so, the word would be sampled according to θ_B. (2) If the author decided that the word should not be a non-informative word, but a content word, the author would then further decide which of the k subtopics this word should be used to describe. To make this decision, the author could use either a document-specific theme coverage distribution ($p(\theta_j|d)$) or a shared theme coverage distribution of all the articles with the same spatiotemporal context as this article ($p(\theta_j|t, l)$). (3) Suppose the j-th subtopic is picked in step (2), the remaining task is simply to sample the word according to θ_j.

We can fit such a spatiotemporal theme model to our weblog data to obtain an estimate of all the parameters in a way similar to the previous work on using mixture models for text mining. The model parameters can then be used to compute various kinds of STTPs.

We now formally present the spatiotemporal theme model.

3.2 The Spatiotemporal Theme Model

Let $C = \{(d_1, \tilde{t}_1, \tilde{l}_1), (d_2, \tilde{t}_2, \tilde{l}_2), ..., (d_n, \tilde{t}_n, \tilde{l}_n)\}$ be a weblog collection where $\tilde{t}_i \in T = \{t_1, t_2, ..., t_{|T|}\}$ is a time stamp and $\tilde{l}_i \in L = \{l_1, l_2, ..., l_{|L|}\}$ is a location label. Suppose $\Theta = \{\theta_1, ..., \theta_k\}$ are k global themes. The likelihood of a word w in document d of time t and location l according to our mixture model is

$$p(w : d, t, l) = \lambda_B p(w|\theta_B) + (1 - \lambda_B) \sum_{j=1}^{k} p(w, \theta_j|d, t, l)$$

where λ_B is the probability of choosing θ_B.

We may decompose $p(w, \theta_j|d, t, l)$ in different ways to obtain interesting special cases of this general mixture model; some of them will be discussed later in Section 3.5. For spatiotemporal analysis of themes, we decompose it as follows:

$$\begin{aligned} p(w, \theta_j|d, t, l) &= p(w|\theta_j)p(\theta_j|d, t, l) \\ &= p(w|\theta_j)((1 - \lambda_{TL})p(\theta_j|d) + \lambda_{TL}p(\theta_j|t, l)) \end{aligned}$$

$$
\begin{aligned}
p(z_{d,w}=j) &= \frac{(1-\lambda_B)p^{(m)}(w|\theta_j)[(1-\lambda_{TL})p^{(m)}(\theta_j|d)+\lambda_{TL}p^{(m)}(\theta_j|t_d,l_d)]}{\lambda_B p(w|B)+(1-\lambda_B)\sum_{j'=1}^{k}p^{(m)}(w|\theta_{j'})[(1-\lambda_{TL})p^{(m)}(\theta_{j'}|d)+\lambda_{TL}p^{(m)}(\theta_{j'}|t_d,l_d)]}\\
p(y_{d,w,j}=1) &= \frac{\lambda_{TL}p^{(m)}(\theta_j|t_d,l_d)}{(1-\lambda_{TL})p^{(m)}(\theta_j|d)+\lambda_{TL}p^{(m)}(\theta_j|t_d,l_d)}\\
p^{(m+1)}(\theta_j|d) &= \frac{\sum_{w\in V}c(w,d)p(z_{d,w}=j)(1-p(y_{d,w,j}=1))}{\sum_{j'=1}^{k}\sum_{w\in V}c(w,d)p(z_{d,w}=j')(1-p(y_{d,w,j'}=1))}\\
p^{(m+1)}(\theta_j|t,l) &= \frac{\sum_{d:t_d=t,l_d=l}\sum_{w\in V}c(w,d)p(z_{d,w}=j)p(y_{d,w,j}=1)}{\sum_{d:t_d=t,l_d=l}\sum_{j'=1}^{k}\sum_{w\in V}c(w,d)p(z_{d,w}=j')p(y_{d,w,j'}=1)}\\
p^{(m+1)}(w|\theta_j) &= \frac{\sum_{d\in C}c(w,d)p(z_{d,w}=j)}{\sum_{w'\in V}\sum_{d\in C}c(w',d)p(z_{d,w'}=j))}
\end{aligned}
$$

Figure 2: EM updating formulas for the spatiotemporal theme model

where λ_{TL} is a parameter to indicate the probability of using the theme coverage distribution of the spatiotemporal context to choose a theme. $p(w|\theta_j)$ gives us the word distribution for each theme, whereas $p(\theta_j|t,l)$ gives a time and location-specific distribution of the themes, which we could exploit to compute various kinds of STTPs. The log likelihood of the whole collection C is thus

$$
\begin{aligned}
\log p(C) &= \sum_{d\in C}\sum_{w\in V}c(w,d)\times\log[\lambda_B p(w|\theta_B)\\
&+(1-\lambda_B)\sum_{j=1}^{k}p(w|\theta_j)((1-\lambda_{TL})p(\theta_j|d)+\lambda_{TL}p(\theta_j|t_d,l_d))]
\end{aligned}
$$

where $c(w,d)$ is the count of word w in document d, t_d and l_d are the time and location labels of d.

3.3 Parameter Estimation

In this section, we discuss how we estimate the parameters of the spatiotemporal theme model above using the maximum likelihood estimator, which chooses parameter values to maximize the data likelihood.

The general model has many parameters to estimate. However, for the purpose of spatiotemporal weblog mining, we will regularize the model by fixing some parameters. First, we set the background model as follows:

$$p(w|\theta_B)=\frac{\sum_{d\in C}c(w,d)}{\sum_{w\in V}\sum_{d\in C}c(w,d)}$$

Second, we set λ_B and λ_{TL} manually, which we will further discuss later in Section 4.

The parameters remaining to be estimated are thus reduced to: (1) the global theme models, $\Theta=\{\theta_1,...\theta_k\}$; (2) the document theme probabilities $p(\theta_j|d)$ where $1\le j\le k, d\in C$; and (3) the spatiotemporal theme probability $p(\theta_j|t,l)$ where $1\le j\le k, t\in T, l\in L$.

We may use the Expectation-Maximization (EM) algorithm [4] to estimate all these parameters by maximizing the data likelihood. The updating formulas are shown in Figure 2. In these formulas, $\{z_{d,w}\}$ is a hidden variable and $p(z_{d,w}=j)$ indicates the probability that word w in document d is generated using theme j. $\{y_{d,w,j}\}$ is another hidden variable and $p(y_{d,w,j}=1)$ indicates the probability that word w is generated using theme θ_j, and θ_j has been chosen according to the spatiotemporal theme coverage distribution ($p(\theta_j|t,l)$), as opposed to the document-specific theme distribution ($p(\theta_j|d)$).

The EM algorithm will terminate when it achieves a local maximum of the log likelihood. It may not reach the global optimal solution when there are multiple maximums. In our experiments, we use multiple trials to improve the local maximum we obtain.

The time complexity of each EM iteration is $O(knu+k|T||L||V|)$, where u is the average number of **unique** words in a document.

3.4 Spatiotemporal Pattern Analysis

Once we have all the parameters estimated using the EM algorithm, we may compute various kinds of STTP patterns using these parameters.

The theme life cycle for a given location $\tilde{l}$ can be obtained by computing

$$p(t|\theta_j,\tilde{l})=\frac{p(\theta_j|t,\tilde{l})p(t,\tilde{l})}{\sum_{\tilde{t}\in T}p(\theta_j|\tilde{t},\tilde{l})p(\tilde{t},\tilde{l})}$$

where $p(t,\tilde{l})$ is given by the word count in time period t at location $\tilde{l}$ divided by the total word count in the collection.

The theme snapshot given time stamp $\tilde{t}$ can be obtained by computing

$$p(\theta_j,l|\tilde{t})=\frac{p(\theta_j|\tilde{t},l)p(\tilde{t},l)}{\sum_{\tilde{l}\in L,}\sum_{j=1}^{k}p(\theta_j|\tilde{t},\tilde{l})p(\tilde{t},\tilde{l})}$$

With the theme life cycles and theme snapshots, various spatiotemporal patterns can be discovered and analyzed. For example, theme shifting can be analyzed by plotting the life cycles of the same theme over different locations together, while theme spreading can be discovered by comparing the theme snapshots of consecutive time periods. We will show examples of such spatiotemporal pattern analysis in Section 4.

3.5 Generality of the Model

Although motivated by specific needs in blog mining, the probabilistic model we proposed is quite general. In this section, we show that several existing models can be viewed as special cases of the spatiotemporal model when we make different simplification assumptions about $p(w,\theta_j|d,t,l)$.

First, if we assume that the choice of a theme and the generation of a word do not depend on the time and location, $p(w,\theta_j|d,t,l)$ will be simplified as $p(w,\theta_j|d)$, which is $p(w|\theta_j,d)p(\theta_j|d)$. Integrated with the assumption $p(w|\theta_j,d)=p(w|\theta_j)$ (generating a word from a given theme does not depend on a specific document), we will obtain the following

simple mixture model:

$$p(w:d) = \lambda_B p(w|\theta_B) + (1-\lambda_B)\sum_{j=1}^{k}[p(w|\theta_j)p(\theta_j|d)]$$

This is exactly the flat mixture model used in [27] and [21]. Note that setting $\lambda_{TL} = 0$ would also lead to this flat model. If we further drop the background model, this model will be equivalent to the PLSI model discussed in [12].

Another possible assumption is that the generation of a document only depends on time but not on location. Correspondingly, we have $p(\theta_j|t,l) = p(\theta_j|t)$. This leads to the following temporal model:

$$p(w:d,t) = \lambda_B p(w|\theta_B) + (1-\lambda_B)\sum_{j=1}^{k} p(w|\theta_j)[(1-\lambda_T)p(\theta_j|d) + \lambda_T p(\theta_j|t_d)]$$

This temporal theme model, although not discussed in existing work, is able to extract location-independent theme life cycles, similar to the life cycle patterns studied in [21]. Indeed, the theme life cycle can be obtained as

$$p(t|\theta_j) = \frac{p(\theta_j|t)p(t)}{\sum_{\tilde{t}\in T} p(\theta_j|\tilde{t})p(\tilde{t})}$$

.

4. EXPERIMENTS AND RESULTS

We apply the proposed spatiotemporal theme model to three different data sets. For each data set, we extract a number of most salient themes and analyze their life cycles and theme snapshots. Experiments show that the proposed model performs well for different types of topics and can reveal interesting spatiotemporal patterns in weblogs.

4.1 Data Set Construction

As discussed in Section 1, themes are subtopics associated with a broad event (or topic). Therefore, we construct each data set by collecting blog entries that are relevant to a given topic. Note that the proposed method could be applied to any collection with time and location information.

We select three topics and construct a data set for each one by submitting a time-bounded query to Google Blog Search[1] and collecting the blog entries returned. Each entry has a pointer to the page containing its author profile. For privacy concerns, we only keep the location information. Since schema matching from different blog providers is not our focus, we only collect the blog entries from MSN Space. The basic information of each data set is presented in Table 3:

Data Set	# docs	Time Span(2005)	Query
Katrina	9377	08/16 - 10/04	Hurricane Katrina
Rita	1754	08/16 - 10/04	Hurricane Rita
iPod Nano	1720	09/02 - 10/26	iPod Nano

Table 3: Basic information of three data sets

We extract free text contents, time stamps and location labels from each document. Krovetz stemmer [15] is used to stem the text. We intentionally did not perform stop word pruning in order to test the robustness of the model.

[1]http://blogsearch.google.com

Originally, the smallest unit of a time stamp is a *day* and the smallest granularity of a location is a *city*. We group the time stamps and locations so that the data in each active unit (t,l) will not be too sparse. (Exactly how to group them will be discussed later in this section.) We then build an index for each data set with Lemur Toolkit[2], on top of which the proposed theme model is implemented.

For each data set, we design our experiments as follows: (1) group the time stamps and locations into appropriate units; (2) apply the spatiotemporal model to extract a number of salient themes; (3) with the parameters estimated, compute the life cycle for each theme at each location and compute the theme snapshot for each time period; (4) visualize the results with life cycle plots and snapshot maps. The experiment details and results are discussed below.

4.2 Parameter Setting

In the spatiotemporal theme model, there are several user-input parameters which provide flexibility for the spatiotemporal theme analysis. These parameters are set empirically. In principle, it is not easy to optimize these parameters without relying on domain knowledge and information about the goal of the data analysis. However, the nature of this mining task is to provide user flexibility to explore the spatiotemporal text data with their belief about the data. We expect that the change of these parameters will not affect the major themes and trends but provide flexibility on analyzing them. The effect of the parameters is as follows.

Generally, we expect each discovered theme to be semantically coherent and distinctive from the general information of the collection, which is captured by the background model. λ_B controls the strength of the background model, and should be set based on how discriminative we would like the extracted themes to be. In practice, a larger λ_B would cause the stop words to be automatically excluded from the top probability words in each theme language model. However, an extremely large λ_B could attract too much useful information into background and make the component theme difficult to interpret. Empirically, a suitable λ_B for blog documents can be chosen between 0.9 and 0.95.

λ_{TL} controls the modeling of spatiotemporal theme distributions. A higher λ_{TL} would allow more content information of a document to be used to learn the spatiotemporal theme distribution, leaving little room for variation in individual documents. $\lambda_{TL} = 1$ would essentially pool all the documents of the same time and location and force them to use the same spatiotemporal theme distribution, whereas $\lambda_{TL} = 0$ would cause the spatiotemporal theme model to degenerate to the flat theme model. Empirically, a good selection of λ_{TL} lies between 0.5 and 0.7.

Parameter k represents the number of subtopics in a collection which can be set based on any prior knowledge about the event. When no domain knowledge is available as in our experiments, we follow [21] to determine the number of themes by enumerating multiple possible values of k and drop the themes with a significant low value of $\frac{1}{|C|}\sum_{d\in C} p(\theta|d)$.

Note that the granularity of time is also a parameter which should be set carefully. A too coarse time granularity may miss interesting bursting patterns. Meanwhile, a too small granularity makes the information in each time interval sparse, which causes the life cycle plots to be sen-

[2]http://www.lemurproject.org/

Theme 1	Theme 2	Theme 3	Theme 4	Theme 5	Theme 6
Government Response	**New Orleans**	**Oil Price**	**Praying and Blessing**	**Aid and Donation**	**Personal Life**
bush 0.0716374	city 0.0633899	price 0.0772064	god 0.141807	donate 0.120228	i 0.405526
president 0.0610942	orleans 0.0540974	oil 0.0643189	pray 0.047029	relief 0.0769788	my 0.11688
federal 0.0514114	new 0.034188	gas 0.0453731	prayer 0.0417175	red 0.0702266	me 0.0601333
govern 0.0476977	louisiana 0.0234546	increase 0.0209058	love 0.0307544	cross 0.0651472	am 0.0291511
fema 0.0474692	flood 0.0227215	product 0.0202912	life 0.025797	help 0.0507348	think 0.0150206
administrate 0.0233903	evacuate 0.0211225	fuel 0.0188067	bless 0.025475	victim 0.0360877	feel 0.0123928
response 0.0208351	storm 0.01771328	company 0.0181833	lord 0.0177097	organize 0.0220194	know 0.0114889
brown 0.0199573	resident 0.0168828	energy 0.0179985	jesus 0.0162096	effort 0.0207279	something 0.00774544
blame 0.0170033	center 0.0165427	market 0.0167884	will 0.0139161	fund 0.0195033	guess 0.00748368
governor 0.0142153	rescue 0.0128347	gasoline 0.0123526	faith 0.0120621	volunteer 0.0194967	myself 0.00687533

Table 1: Selected themes extracted from Hurricane Katrina data set

Theme 1	Theme 2	Theme 3	Theme 4	Theme 5	Theme 6
Personal Life	**New Orleans**	**Government**	**Oil Price**	**Storm in Texas**	**Cause of Disaster**
i 0.378907	orleans 0.0877966	bush 0.0799938	oil 0.0929594	texas 0.0827817	warm 0.0408111
my 0.130939	new 0.0863562	president 0.0380268	price 0.0910098	storm 0.0677999	global 0.0362097
me 0.052442	city 0.0448588	govern 0.0336889	gas 0.08018	wind 0.0457836	cancer 0.0338968
am 0.0296791	levee 0.0310737	disaster 0.0327676	market 0.0274118	galveston 0.0398179	climate 0.0310257
her 0.0242092	water 0.0285081	federal 0.0291178	product 0.0247364	houston 0.0374735	change 0.0305301
feel 0.0133843	black 0.0242845	war 0.0283924	company 0.0239605	coast 0.0357728	rise 0.015924
friend 0.0114668	flood 0.0233388	iraq 0.0245395	energy 0.0235243	evacuate 0.0266514	surface 0.0123937
work 0.0102118	police 0.0175093	agent 0.0172738	cent 0.0216559	resident 0.0172511	scientist 0.0123253
love 0.00720455	superdome 0.0107274	katrina 0.0138582	barrel 0.0192692	landfall 0.0159967	temperature 0.0105363
life 0.00715712	neighborhood 0.00882865	response 0.0120409	refinery 0.0175738	tropic 0.012126	ocean 0.0102157

Table 2: Selected themes extracted from Hurricane Rita data set

sitive and with sharp variations. We address this problem by first selecting a reasonable small time granularity (1 day) and use a sliding window to smooth the theme life cycle at a later stage. For example, we may use $\tilde{p}(t_i|\theta, l) = \frac{1}{3}[p(t_{i-1}|\theta, l) + p(t_i|\theta, l) + p(t_{i+1}|\theta, l)]$ to substitute $p(t_i|\theta, l)$ when plotting.

In the following sections, we present interesting themes, theme life cycles and theme snapshots discovered from the three data sets.

4.3 Hurricane Katrina Data Set

The Hurricane Katrina data set is the largest one in our experiments. 7118 documents out of 9377 have location information. We vary the time granularity from a day to a week. The extracted themes are not sensitive to this granularity change. We set the granularity of location as a *state* and analyze the theme snapshot within the United States.

The most salient themes extracted from the Hurricane Katrina data set are presented in Table 1, where we show the top probability words of each theme language model. The semantic labels of each theme are presented in the second row of Table 1. We manually label each theme with the help of the documents with highest $p(\theta|d)$. A few less meaningful themes are dropped as noise.

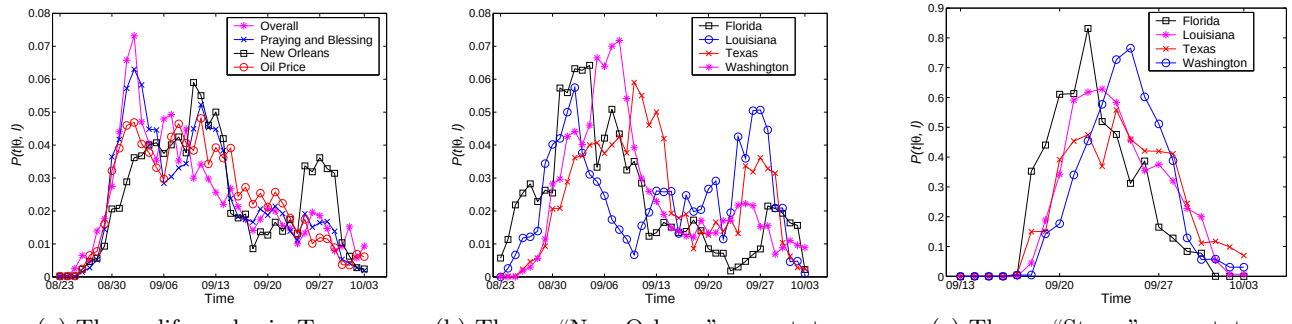
From Table 1, we can tell that theme 1 suggests the concern about "Government Response" to the disaster; theme 2 discusses the subtopic related to "New Orleans"; theme 3 represents people's concern about the increase of "Oil Price"; theme 4 is about "praying and blessing" for the victims; and theme 5 covers the aid and donations made for victims. Unlike theme 1 to theme 5, the semantics of which can be inferred from the top probability words, theme 6 is hard to interpret directly from the top words. By linking back to the original documents, we find that the documents with highest probability $p(\theta|d)$ for theme 6 tend to talk about personal life and experiences of the author. This is interesting and reasonable because weblogs are associated with personal contents. Indeed, we observe that a similar theme also occurs in other two data sets.

We then plot and compare the theme life cycles at the same location and life cycles of a theme over states. Interesting results are selectively shown in Figure 3(a) and (b).

Figure 3(a) shows the life cycles of different themes in Texas. The "Overall" life cycle shows the coverage of the overall topic measured by the collection size over time. Clearly, all themes grow rapidly during the first week, in which Hurricane Katrina was active. The theme "praying and blessing" starts dropping after ten days and increases again during the third week. In the same week, the discussion about "New Orleans" reaches the peak. Comparing with real time line of Hurricane Katrina, we find that this is the week that the mayor of New Orleans ordered evacuation. The theme "New Orleans" rises again around late September. The public concern of "Oil Price", however, shows a stable high probability until the fourth week.

Figure 3(b) plots the life cycles of theme "New Orleans" at different states. We observe that this theme reaches the highest probability first in Florida and Louisiana, followed by Washington and Texas, consecutively. During early September, this theme drops significantly in Louisiana while still strong in other states. We suppose this is because of the evacuation in Louisiana. Surprisingly, around late September, an arising pattern can be observed in most states, which is most significant in Louisiana. Since this is the time period in which Hurricane Rita arrived, we surmise that Hurricane Rita has an impact on the discussion of Hurricane Katrina. This is reasonable since people are likely to mention the two hurricanes together or make comparisons. We can find more clues to this hypothesis from Hurricane Rita data set.

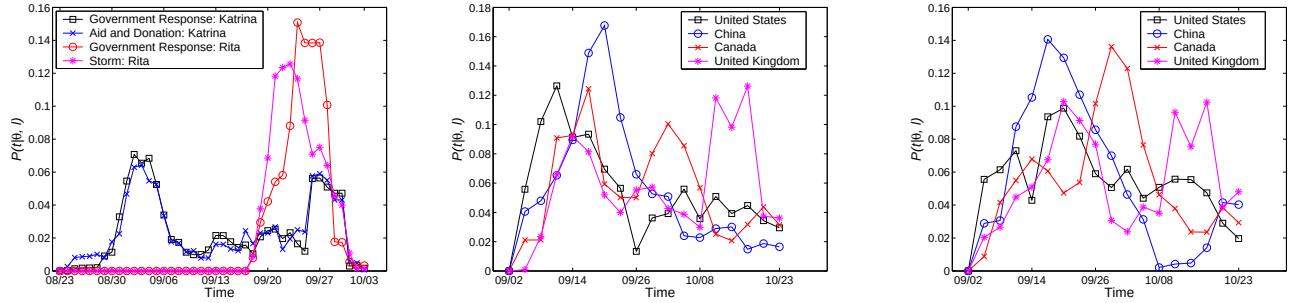
Representative snapshots for theme "Government Response" over five weeks are presented in Figure 4. The darker the color is, the larger the $p(\theta, l|t)$ is, but the color cannot be compared across the snapshots because $p(\theta, l|t)$ is conditioned on the time t, which differs in each snapshot. From Figure 4, we observe that at the first week of Hurricane Katrina, the theme "Government Response" is the strongest in the southeast states, especially those along the Gulf of Mexico. In week 2, we can clearly see the pattern that the theme is spreading towards the north and western states. This pattern continues over week 3, in which the theme is distributed much more uniformly over the States. However, in week 4, we observe that the theme converges to east states and southeast coast again. Interestingly, this week happens



(a) Theme life cycles in Texas (Hurricane Katrina)

(b) Theme "New Orleans" over states (Hurricane Katrina)

(c) Theme "Storm" over states (Hurricane Rita)



(d)Theme life cycles in Louisiana (Hurricane Rita)

(e) Theme "Release of Nano" over states (iPod Nano)

(f) Theme "Nano Features" over states (iPod Nano)

Figure 3: Theme life cycle patterns from three data sets

to overlap with the first week of Hurricane Rita, which may raise the public concern about government response again in those areas. In week 5, the theme becomes weak in most inland states and most of the remaining discussions are along the coasts. We will see comparable patterns at the same time periods from Hurricane Rita data set. Another interesting observation is that this theme is dramatically weakened in Louisiana during week 2 and 3, and becomes strong again from the fourth week. Week 2 and 3 are consistent with the time of evacuation in Louisiana.

4.4 Hurricane Rita Data Set

In the Hurricane Rita data set, 1403 documents out of 1754 have location labels. As in Hurricane Katrina, we choose a *state* as the smallest granularity of locations.

The most salient themes extracted from Hurricane Rita data set are shown in Table 2.

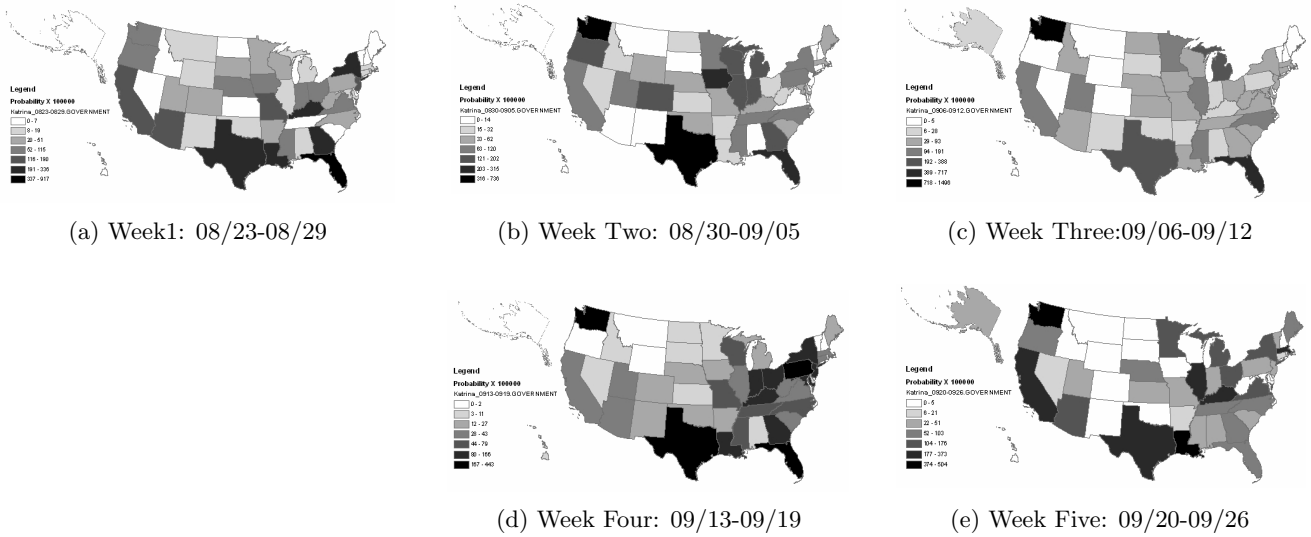
Compared with the themes extracted from the Hurricane Katrina data set, we observe that the two data sets share several similar themes. Besides the "Personal Life" theme, we see that the theme "New Orleans", "Government Response", and "Oil Price" occur in both collections. This is reasonable because the two events are comparable. The Hurricane Rita data set, however, has its specific themes such as "the Storm in Texas" and "Cause of the Disaster". We notice that the theme "Government Response" of Hurricane Rita covers extra politics contents such as Iraq War. Some themes of Hurricane Katrina, such as "Praying" and "Donation", do not appear to be salient in Hurricane Rita.

The interesting theme life cycles of Hurricane Rita are presented in Figure 3 (c) and (d). Figure 3(c) shows the life cycles of theme "Storm" over different states. Similar to Hurricane Katrina, at the very beginning, Florida is the most active state, which is also the first state where the theme reaches its peak. Shortly after Florida, the theme becomes the strongest in Louisiana, followed by Texas and Washington. In most states, the theme life cycle drops monotonically after the peak. All the life cycles fade out within two weeks from 9/17, which indicates that the impact of Hurricane Rita may not be as high as Hurricane Katrina.

Figure 3(d) compares the theme life cycles in Louisiana between Hurricane Katrina and Hurricane Rita. The two Hurricane Katrina themes share similar life cycle patterns and so are the two themes of Hurricane Rita. In Louisiana, the discussion of Hurricane Katrina drops rapidly around early September and rises again significantly at the last week of September. Interestingly, this arising is just shortly after the significant rising of the discussion about Hurricane Rita. This further strengthens our hypothesis that the two events have interactive impacts in weblogs. Indeed, nearly 40% blog entries which mentioned Hurricane Rita after September 26th also mentioned Hurricane Katrina.

The theme snapshots over the first two weeks of Hurricane Rita show that the discussion of Hurricane Rita did not spread so significantly as the first two weeks of Hurricane Katrina. Instead, the spatial patterns are similar to the last two weeks of Hurricane Katrina, which are roughly around the same time period. We present the snapshots of the theme "Oil Price" in Figure 5.

During the first week of Hurricane Rita, we observe that the theme "Oil Price" is already widespread over the States. In the following week, the topic does not further spread; instead, it converges back to the states strongly affected by the hurricane. Comparable patterns for the same theme can be found during the last two weeks of Hurricane Katrina. This further implies that the two comparable events have interacting impact on the public concerns about them.



(a) Week1: 08/23-08/29 (b) Week Two: 08/30-09/05 (c) Week Three:09/06-09/12

(d) Week Four: 09/13-09/19 (e) Week Five: 09/20-09/26

Figure 4: Snapshots for theme "Government Response" over the first five weeks of Hurricane Katrina

4.5 iPod Nano Data Set

This data set contains 1387 documents with location labels. We assume that the theme patterns have more significant difference between different countries rather than between states inside United States. Therefore, we discretize locations into *countries*.

Table 4 shows the interesting themes extracted from this data set. Again, we observe the personal life theme, which may be a characteristic theme for weblogs.

Theme 1	Theme 2	Theme 3	Theme 4
Release of Nano	**Marketing**	**Special Features**	**Personal Life**
ipod 0.2875	will 0.0306	your 0.0598	i 0.2489
nano 0.1646	market 0.0273	1 0.0478	you 0.0627
apple 0.0813	search 0.0270	music 0.0455	have 0.0312
september 0.0510	apple 0.0257	song 0.0378	my 0.0269
mini 0.0442	company 0.0215	display 0.0209	am 0.0220
screen 0.0242	itune 0.0200	shrink 0.0182	know 0.0214
new 0.0200	phone 0.0199	4gb 0.0130	want 0.0148
mp3 0.0155	web 0.0186	color 0.0102	thing 0.0146
thin 0.0140	microsoft 0.0185	pencil-thin 0.0100	would 0.0130
shuffle 0.0127	motorola 0.0151	model 0.0096	think 0.0110

Table 4: Selected themes from iPod Nano data set

In the rest three themes, theme 1 is about the news that iPod Nano was introduced. Theme 2 is about the marketing of Apple and how it is related to other business providers. Theme 3 is about specific features of iPod Nano.

We compare the life cycles of themes over different countries. Our expectation is that the life cycle of themes in the United States would evolve faster than those outside. The results are shown in Figure 3 (e) and (f).

For the theme about the release of iPod Nano, United States is indeed the first country where it reaches the top of its life cycle, followed by Canada, China, and United Kingdom consecutively. The theme in China presents a sharp growing and dropping, which indicates that most discussions there are within a short time period. The life cycles in Canada and United Kingdom both have two peaks.

Similar patterns can be found in the theme discussing the specific features of Nano. All life cycles start around early September. At the very beginning, discussions in the United States surge more quickly than in any other countries. The theme reaches its peak in Canada posteriorly to the other countries. There are also two peaks in United Kingdom.

To summarize, the experiments on three different data sets show that the spatiotemporal theme model we proposed in Section 3 can extract themes and their spatiotemporal patterns effectively. The comparative analysis of theme life cycles and theme snapshots is potentially useful to reveal interesting patterns and to answer a lot of questions.

5. RELATED WORK

To the best of our knowledge, the problem of spatiotemporal text mining has not been well studied in existing work.

Most existing text mining work (e.g., [22, 21]) does not consider the temporal and location context of text. Li and others proposed a probabilistic model to detect retrospective news events by explaining the generation of "four Ws[3]" from each news article [18]. However, their work considers time and location as independent variables, and aims at discovering the reoccurring peaks of events rather than extract the spatiotemporal patterns of themes.

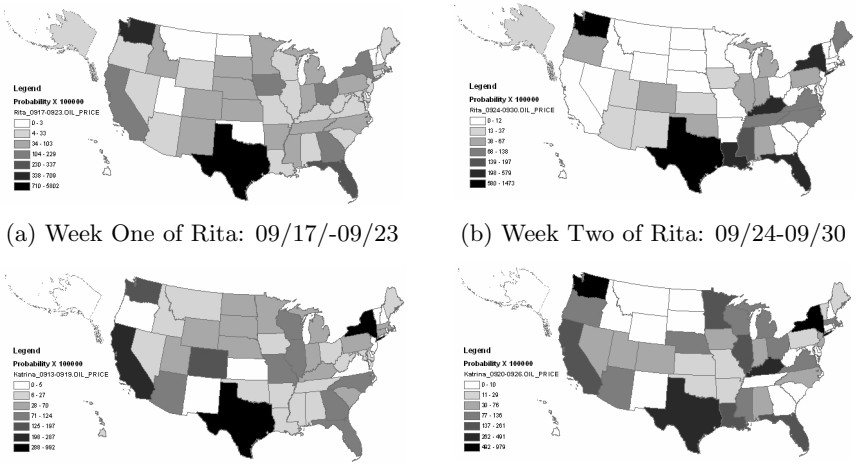
Some other related work can be summarized in the following several lines.

5.1 Text Mining

Text clustering is a well studied problem relevant to our work. Some previous studies have presented several probabilistic models to model themes in documents (e.g., PLSI [12] and LDA [1]). In [27], PLSI is extended to include a background component to explain the non-informative background words, which was also adopted in [21]. A cross-collection mixture model was proposed in [27] to support comparative text mining. However, none of these models takes into account the temporal or spatial information.

Temporal text mining has been addressed recently. Kleinberg's work discovers bursty and hierarchical structures in streams [13] by converting text streams to temporal frequency data and using an infinite-state automaton to model

[3]who (persons), when (time), where (locations) and what (keywords)



(a) Week One of Rita: 09/17/-09/23 (b) Week Two of Rita: 09/24-09/30

(c) Week Four of Katrina: 09/13-09/19 (d) Week Five of Katrina: 09/20-09/26

Figure 5: Snapshots for theme "Oil Price" of the first two weeks of Hurricane Rita

the stream. [6] proposed a method to identify information novelty in news stories, which can be applied to model content evolvement over time. [21] proposed a probabilistic approach to discovering evolutionary theme patterns, which first extracts themes with the flat theme model and then segments the whole collection with a Hidden Markov Model. The life cycle of each theme is plotted as the strength of the theme over time. This approach, however, did not provide a unified probabilistic model for theme extraction and theme life cycles. [9] presents a generative model similar to the one in [1] to extract scientific topics from PNAS abstracts and a post hoc analysis on the popularity of topics to detect the hot and cold topics. Our work differs from theirs in that we model the temporal dynamics of themes simultaneously with theme extraction in the statistical model. Another related work to theme life cycle analysis is [24], where a Multinomial PCA model is used to extract themes from text and analyze temporal trends. However, none of this work models the spatial information of a text collection.

5.2 Weblog Analysis

Another line of research related to our work is weblog analysis and mining. Existing work has explored either structural analysis on communities [26, 16] and temporal analysis on blog contents [8, 11]. Our work differs from the existing work in two aspects: (1) we model the multiple themes within each blog article; (2) we correlate the contents, location and time of articles in a unified probabilistic model. None of these has been done in the previous work of weblog analysis.

Kumer and others showed that the structure and interest clusters on blogspace are highly correlated to the locality property of weblogs [17]. Although this work considered temporal and spatial *distribution* of weblogs, neither this work nor any other previous work has addressed the *content* analysis with spatiotemporal information. We consider this work as an important evidence that spatial analysis on weblog content is desired.

Some existing work further explored content and structure evolutions of weblogs for higher level tasks. For example, Gruhl and others' work in 2004 modeled information diffusion through blogspage by categorizing temporal topic patterns into spikes and chatter [11]. Their following work in 2005 explored the spike patterns of discussion of books to predict spikes in their sales rank [10].

The general spatiotemporal theme analysis methods proposed in our work can provide fundamental utilities to facilitate such higher-level predictions.

5.3 Others

Spatiotemporal data mining on numerical data and moving objects has been well studied [5, 20]. [23] present a spatiotemporal clustering method to detect the emerging space-time clusters. However, these techniques aim at analyzing explicit data objects, which cannot be used for extracting and analyzing latent theme patterns from a text collection.

Topic detection and tracking [2, 25, 19] aims at detecting emerging new topics and identifying boundaries of existing events. Morinaga and Yamanishi tracked the dynamics of topic trends in real time text stream [22]. They assumed that the dynamics of each topic bears a Gaussian distribution. Trend Detection [14] detects emerging trends of topics from text. However, most of those works focus on "events" rather than themes (i.e. they assume one article belongs to one event while we assume a blog entry can consist of multiple themes). Moreover, this work has not considered spatial patterns of topics.

6. CONCLUSIONS

Weblogs usually have a mixture of subtopics and exhibit spatiotemporal content patterns. Discovering themes and modeling their spatiotemporal patterns are beneficial not only for weblog analysis, but also for many other applications and domains. In this paper, we define the general problem of spatiotemporal theme patterns discovery and propose a novel probabilistic mixture model which explains the generation of themes and spatiotemporal theme patterns simultaneously. With this model, we discover spatiotemporal theme patterns by (1) extracting common themes from weblogs; (2) generating theme life cycles for each location; and

(3) generating theme snapshots for each given time period. Evolution of patterns can be analyzed through comparative analysis of theme life cycles and theme snapshots.

We evaluate our approach on three weblog collections about different events. Experiment results show that the proposed approach can discover interesting themes, theme life cycles and theme snapshots effectively. We show that the proposed probabilistic model is quite general, covering existing theme models as special cases. Therefore, the probabilistic model is generally applicable not only to any text collections with time and location information, but also for other text mining problems. Our approach can serve as a fundamental utility for higher level tasks of analysis and research based on spatiotemporal patterns, such as prediction of user behavior.

There are several interesting directions to extend our work. In this work, we have not considered the adjacency and distance between time and locations. How to model these factors and select an appropriate spatiotemporal granularity is an interesting problem. Modeling the content variations of themes over time and location is also a very interesting direction.

7. ACKNOWLEDGEMENTS

We sincerely thank the anonymous reviewers for their extensive useful comments. We also thank Yang Chen for helping visualize the spatiotemporal theme distributions. This work is in part supported by the National Science Foundation under award numbers 0425852, 0347933, and 0428472.

8. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] S. Boykin and A. Merlino. Machine learning of event segmentation for news on demand. *Commun. ACM*, 43(2):35–41, 2000.

[3] W. B. Croft and J. Lafferty, editors. *Language Modeling and Information Retrieval*. Kluwer Academic Publishers, 2003.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statist. Soc. B*, 39:1–38, 1977.

[5] U. Fayyad, D. Haussler, and P. Stolorz. Mining scientific data. *Commun. ACM*, 39(11):51–57, 1996.

[6] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th International Conference on World Wide Web*, pages 482–490, 2004.

[7] K. E. Gill. Blogging, rss and the information landscape: A look at online news. In *WWW 2005 Workshop on the Weblogging Ecosystem*, 2005.

[8] N. Glance, M. Hurst, and T. Tornkiyo. Blogpulse: Automated trend discovery for weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[9] T. L. Griffiths and M. Steyvers. Fiding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl.1):5228–5235, 2004.

[10] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proceedings of KDD '05*, pages 78–87, 2005.

[11] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, pages 491–501, 2004.

[12] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of SIGIR '99*, pages 50–57, 1999.

[13] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of KDD '02*, pages 91–101, 2002.

[14] A. Kontostathis, L. Galitsky, W. M. Pottenger, S. Roy, and D. J. Phelps. A survey of emerging trend detection in textual data mining. *Survey of Text Mining*, pages 185–224, 2003.

[15] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of SIGIR '93*, pages 191–202, 1993.

[16] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proceedings of the 12th International Conference on World Wide Web*, pages 568–576, 2003.

[17] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *Commun. ACM*, 47(12):35–39, 2004.

[18] Z. Li, B. Wang, M. Li, and W.-Y. Ma. A probabilistic model for retrospective news event detection. In *Proceedings of SIGIR '05*, pages 106–113, 2005.

[19] J. Ma and S. Perkins. Online novelty detection on temporal sequences. In *Proceedings of KDD '03*, pages 613–618, 2003.

[20] N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. W. Cheung. Mining, indexing, and querying historical spatiotemporal data. In *Proceedings of KDD '04*, pages 236–245, 2004.

[21] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of KDD '05*, pages 198–207, 2005.

[22] S. Morinaga and K. Yamanishi. Tracking dynamics of topic trends using a finite mixture model. In *Proceedings of KDD '04*, pages 811–816, 2004.

[23] D. B. Neill, A. W. Moore, M. Sabhnani, and K. Daniel. Detection of emerging space-time clusters. In *Proceedings of KDD '05*, pages 218–227, 2005.

[24] J. Perkio, W. Buntine, and S. Perttu. Exploring independent trends in a topic-based search engine. In *Proceedings of WI'04*, pages 664–668, 2004.

[25] K. Rajaraman and A.-H. Tan. Topic detection, tracking, and trend analysis using self-organizing neural networks. In *PAKDD*, pages 102–107, 2001.

[26] B. Tseng, J. Tatemura, and Y. Wu. Tomographic clustering to visualize blog communities as mountain views. In *WWW 2005 Workshop on the Weblogging Ecosystem*, 2005.

[27] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of KDD '04*, pages 743–748, 2004.