# Mining Correlated Bursty Topic Patterns from Coordinated Text Streams

Xuanhui Wang, ChengXiang Zhai, Xiao Hu, Richard Sproat

University of Illinois at Urbana-Champaign
Urbana, IL 61801
{xwang20, czhai, xiaohu, rws}@uiuc.edu

## ABSTRACT

Previous work on text mining has almost exclusively focused on a *single* stream. However, we often have available multiple text streams indexed by the same set of time points (called *coordinated text streams*), which offer new opportunities for text mining. For example, when a major event happens, all the news articles published by different agencies in different languages tend to cover the same event for a certain period, exhibiting a *correlated bursty topic pattern* in all the news article streams. In general, mining correlated bursty topic patterns from coordinated text streams can reveal interesting latent associations or events behind these streams. In this paper, we define and study this novel text mining problem. We propose a general probabilistic algorithm which can effectively discover correlated bursty patterns and their bursty periods across text streams even if the streams have *completely different vocabularies* (e.g., English vs Chinese). Evaluation of the proposed method on a news data set and a literature data set shows that it can effectively discover quite meaningful topic patterns from both data sets: the patterns discovered from the news data set accurately reveal the major common events covered in the two streams of news articles (in English and Chinese, respectively), while the patterns discovered from two database publication streams match well with the major research paradigm shifts in database research. Since the proposed method is general and does not require the streams to share vocabulary, it can be applied to any coordinated text streams to discover correlated topic patterns that burst in multiple streams in the same period.

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

**General Terms:** Algorithms

**Keywords:** Correlated bursty patterns, coordinated streams, clustering, reinforcement.


*KDD'07,* August 12–15, 2007, San Jose, California, USA.


## 1. INTRODUCTION

Text streams are ubiquitous and are often naturally formed as new information is incrementally created and accumulated. For example, newswires publish news articles everyday on the Web to report new events to users, generating news streams in different languages such as English, Chinese, and Spanish. Search engines accept and answer end users' queries from all over the world continuously, creating streams of queries. Researchers publish scientific papers year by year, forming literature streams. Blog authors regularly publish blog articles, forming a dynamic stream of blog articles.

One interesting characteristic of a text stream is that there is often an intensive coverage of some topic within a certain period, which we refer to as a *bursty topic pattern*. For example, when a major event happens in the world, all news articles tend to have intensive coverage of the event; as a result, there would be a coverage burst of the topic lasting for a certain period. Similarly, when a new research direction is opened up in a research field, many publications in the new direction tend to be generated, again forming a bursty pattern about the research topic. Mining such bursty topic patterns can help reveal the underlying events and has potentially many applications, such as monitoring opinions, analyzing trends, and summarizing the major topics in a text stream.

So far, text mining research has almost exclusively focused on mining one *single* text stream. For example, the Topic Detection and Tracking (TDT) work [5, 24, 23, 4] has focused on detecting new events and tracking known events in a single news article stream. Other work on extracting bursty patterns has also focused on only one single stream (see, e.g., [18, 19, 12, 8, 15, 14]). However, we often have available multiple related text streams indexed by the same set of time points (called *coordinated text streams*), which offer new opportunities for text mining. In particular, we may discover *correlated bursty topic patterns* from multiple coordinated text streams. A correlated bursty topic pattern refers to simultaneous bursting of some related topics in all the text streams; it is often associated with some underlying event that has influenced the generation of all the text streams involved.

For example, when a major event happens, all the news articles published by different agencies in different languages tend to cover the same event for a certain period, exhibiting a correlated bursty topic pattern in all the news article streams. Exploiting multiple streams to detect the latent

events would be more accurate than using only one single stream as the latter may not be able to distinguish a global event from a local event, leading to mixed mining results. Also, when there is a major shift in research paradigm in a field, all the journals or conferences in the field will likely have a coverage burst of the new research paradigm; again, mining multiple journals or conferences can recover the research paradigm shift more accurately than using only one single publication source.

In general, mining correlated bursty topic patterns from coordinated text streams is quite interesting for several reasons: (1) It can help discover interesting common (causal) events that have influenced all the streams. (2) It can reveal interesting associations and linkages between the involved streams. (3) It can help discover "local" (i.e., stream-specific) patterns more accurately by factoring out the "global noise." For example, identifying correlated bursty topic patterns from news streams in different natural languages such as English and Chinese can not only reveal the same major events covered by both streams, but also create associations of English terms and Chinese terms; such associations would be very useful for cross-lingual information retrieval, integration, and summarization [22, 20]. The discovered common events in multiple news streams can also facilitate discovery of local events specific to one stream (e.g., Chinese news) which would otherwise have be mixed with the common/global events.

In this paper, we define and study the novel problem of mining correlated bursty topic patterns from multiple coordinated text streams. We propose probabilistic mixture models which can identify the bursty patterns and their bursty periods from coordinated streams simultaneously even if the streams have completely different vocabularies (e.g., English and Chinese). The basic idea of our approach is to introduce a latent cause variable to model the underlying events to be discovered and model the text data in multiple streams with a mixture model involving multinomial component topic models. Each topic model is a word distribution with the high probability words indicating the topic content. We do not require the multiple streams to share the same vocabulary; instead, we rely on the correlation between the time distributions of topics to "align" topics from multiple streams. Thus the proposed methods can actually be applied to any discrete data streams, though we have only evaluated it using text streams in this paper. By fitting such a model to the available text streams using the Expectation-Maximization (EM) algorithm, we can obtain the topic models associated with each value of the latent cause variable; these topic models together with their peaking time periods are taken as the correlated bursty topic patterns that we want to discover.

We further propose two extensions to this basic mining approach: (1) We incorporate local dependency (along the time line) into the mixture model to further favor a topic model that can explain well all the documents in a consecutive time period. This allows us to discover consecutive bursty periods. (2) We propose a mutual reinforcement method which allows multiple streams to work together to further improve the quality of the identified correlated bursty patterns by selecting terms that truly have strong global correlations across all the streams.

We test the proposed methods on two data sets – news streams and literature streams. Experiment results show that the proposed methods can effectively discover quite meaningful topic patterns from both data sets: the patterns discovered from the news data set can accurately reveal the major common events covered in the two streams of news articles (in English and Chinese, respectively), while the patterns discovered from two database publication streams match well with the major research paradigm shifts in database research. The proposed two extensions (i.e., local dependency and mutual reinforcement) are also both effective for further improving the quality of discovered patterns.

The rest of the paper is organized as follows. We first review the related work in Section 2. Then we define the problem of mining coordinated streams in Section 3 and present the proposed mining methods in Section 4. We report the experimental results in Section 5 and conclude in Section 6.

## 2. RELATED WORK

Our work is related to several lines of work in text mining, stream data mining, and multilingual natural language processing.

First, the work on Topic Detection and Tracking (TDT) [4, 5, 24, 23] all aims to detect and track events from a stream of news stories, thus is related to our work. However, this body of work all considers a single news stream and does not address the issue of multiple subtopics within a news article. Our work is more related to the *retrospective* version of TDT [24] where the whole stream is analyzed. A main difference between our work and the TDT work is that we consider multiple coordinated text streams and mine correlated bursty topic patterns.

Second, bursty patterns or events are recently studied[18, 19, 12, 8]. In [12], an infinite automaton was proposed to identify bursty features and their bursty structures; it has been used in [13] to identify the bursty evolution of blogspace. However, the work is restricted in only identifying bursty features one by one and does not group the features to find interesting topic patterns. In [18, 19] and [8], bursty features are identified heuristically with multiple steps. For example, in [18, 19], for each named entity and noun phrase in the stream, $\chi^2$ tests are performed to identify the days in which the test scores are higher than a threshold. In [8], binomial distributions are calculated to identify the bursty features. All these methods process the features one by one and only a single stream is analyzed. One problem of such methods is that the results are often quite sensitive to some noisy features which may be incidentally bursty and the bursty pattern is not meaningful. Our method is more robust since it identifies bursty patterns by pooling together many words which share similar patterns. Furthermore, in [18, 19] and [8] it is shown to be difficult for their methods to find long consecutive time periods. In contrast, our methods (especially the local dependency model) can help find long consecutive periods of bursty patterns.

Data streams and time-series data are extensively studied in the database and data mining communities [9, 1]. Much of the emphasis there is on similarity search, which is to find similar time-series sequences given a time-series query (e.g., [3, 21]), and on classification or incremental clustering of data streams (e.g.,[11, 2]).

Temporal information has also been used to identify semantically similar search engine queries [7], to integrate multilingual information [20], and to acquire lexical associations

for transliteration and translation [17]. We propose a mixture model for coordinated text streams with temporal information. It is an extension of Probabilistic Latent Semantic Analysis (PLSA) [10] and is also related to some other recent extensions such as [25, 16].

# 3. PROBLEM FORMULATION

In this section, we formally define the problem of mining correlated bursty topic patterns from multiple coordinated text streams. We first define text stream.

DEFINITION 1 (TEXT STREAM). *A text stream $S$ of length $n$ and with vocabulary $V$ is an ordered sequence of text samples $(S_1, S_2, ..., S_n)$ indexed by time, where $S_i$ is a sequence of words from the vocabulary set $V$ at time point $i$.*

For example, in a news article stream $S$, $S_i$ could be a concatenation of all the news articles published on date $i$, while in a search engine query log stream $S'$, $S'_i$ could be all the queries sent to the search engine on date $i$.

DEFINITION 2 (COORDINATED TEXT STREAMS). *A set of text streams is called coordinated text streams if all the streams share the same time index and have the same length. Formally, a set of $m$ coordinated text streams with length $n$ is $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_m\}$, where $\mathcal{S}_i = (\mathcal{S}_{i1}, ..., \mathcal{S}_{in})$ is the $i$-th stream with vocabulary $V_i$. $\mathcal{S}_{ij}$ is the text sample at time point $j$ in the $i$-th stream, thus it consists of a sequence of words from $V_i$.*

Note that we allow each stream $\mathcal{S}_i$ to have a potentially distinct vocabulary set $V_i$; this allows us to conveniently model text streams in different natural languages such as English, Spanish, and Chinese.

In order to define the concept correlated bursty topic pattern, we first define topic.

DEFINITION 3 (TOPIC). *A topic in stream $\mathcal{S}_i$ is defined as a probability distribution of words in vocabulary set $V_i$. We also call such a word distribution a topic model.*

Using a word distribution (i.e., unigram language model) to represent a topic has been quite common in text mining (see e.g. [10, 6, 25, 16]). Intuitively, a topic model would assign high probabilities to those words that can characterize the topic well. For example, the topic model about the 9–11 terrorist attack may have high probabilities for words such as "attack", "terror", "terrorist", "Afghan", and "Bin Laden", but very small probabilities for words such as "Olympic", "game", "sport", and "swimming", whereas the topic model about an Olympic swimming event would likely be the opposite.

While any topics that we can discover from a text stream would be interesting, we are particularly interested in a special kind of topics which we refer to as "bursty topics." These are topics that are covered intensively within a relatively long consecutive time period in a stream. Bursty topics are interesting because they tend to be associated with some major events, and discovering such events is our goal. We now formally define a bursty topic.

DEFINITION 4 (BURSTY TOPIC). *Let $\theta$ be a topic (model) in stream $\mathcal{S}_i$. Let $t \in [1, n]$ be a time index variable and $p(\theta|t, \mathcal{S}_i)$ be the relative coverage of the topic $\theta$ at time $t$ in stream $\mathcal{S}_i$. $\theta$ is a bursty topic in stream $\mathcal{S}_i$ if $\exists t_1, t_2 \in [1, n]$ such that $t_2 - t_1 \geq \sigma$ and $\forall t \in [t_1, t_2]$, $p(\theta|t, \mathcal{S}_i) \geq \kappa$ where $\sigma$ is a span threshold and $\kappa$ is a coverage threshold. Intuitively, the first condition ensures that the topic is covered in the stream for a relatively long consecutive period; the second ensures that the coverage of the topic is relatively intensive.*

Finally, we define a correlated bursty topic pattern, which is a set of topics (each from a different stream) that are bursty during the same time period. Formally,

DEFINITION 5 (CORRELATED BURSTY TOPIC PATTEN). *A correlated bursty topic pattern in a set of coordinated text streams $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_m\}$ is defined as a set of topics $\{\theta^1, ..., \theta^m\}$ such that $\theta^i$ is a bursty topic in stream $\mathcal{S}_i$ and $\exists t_1, t_2 \in [1, n]$ such that $t_2 - t_1 \geq \sigma$ and $\forall t \in [t_1, t_2]$, $\forall i \in [1, m]$, $p(\theta^i|t, \mathcal{S}_i) \geq \kappa$ where $\sigma$ is a span threshold and $\kappa$ is a coverage threshold.*

According to these definitions, the problem of mining correlated bursty topic patterns from a set of coordinated text streams mainly involves three challenges: (1) We need to discover bursty topics from each stream. As we will show later, a straightforward application of existing approaches to topic discovery cannot effectively detect topics that are bursty. (2) We need to locate the bursty period of a bursty topic. The coverage of a bursty topic may be uneven and unsmooth even during the bursty period, making it a challenge to accurately detect the bursty period boundaries. (3) Since the topics forming a correlated bursty topic pattern must be bursty during the same period in multiple streams, we need to coordinate the discovery of bursty topics in all the streams to more effectively focus on the truly correlated topics. In particular, the discovery of topics in one stream should pay attention to which period other streams suggest to be promising for finding a correlated bursty topic pattern.

# 4. COORDINATED MIXTURE MODEL

In this section, we describe our coordinated mixture model to discover correlated bursty topic patterns from coordinated text streams.

## 4.1 Basic Idea

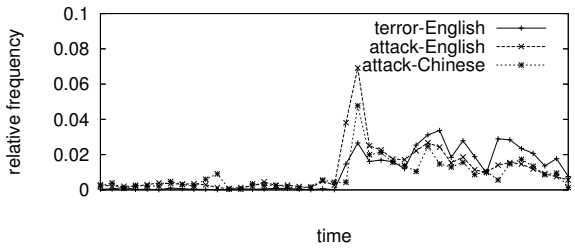



**Figure 1: Examples of bursty words related to 9–11 event in the news data. $x$-axis denotes the time points and $y$-axis is the relative frequency.**

The basic idea of our approach is to align the text samples from different streams based on the shared time stamps and discover topics from multiple streams simultaneously with a single probabilistic mixture model. Recent work such as [10,

6, 16] has shown that probabilistic mixture models are quite effective for discovering topics from text. Their basic idea is to represent a topic by a word distribution and assume that a text collection is "generated" by repeatedly sampling words from a mixture of multiple topic distributions. By fitting the mixture model to the text data, we can then obtain an estimate of each word distribution, which we can take as a discovered topic. Different methods differ in the way of mixing the word distributions and estimating the parameters.

A straightforward way of applying such a method to our problem would be to use a mixture model to discover topics from each stream and then try to match the topics across streams in hope of detecting some topics that happen to burst during the same period. However, there are two problems with this simple approach: (1) We will need to match topics across different streams, which is difficult because the vocabularies of different streams do not necessarily overlap. (2) The topics discovered in each stream may explain the corresponding stream well but not necessarily match the common topics shared by multiple streams. Indeed, a shared topic may not fit a specific stream so well as some variant of the topic.

This analysis suggests that we should somehow make these mixture models designed for different streams "communicate" with each other so that they would all focus more on discovering the common topics shared by all streams. We achieve this goal by aligning the text samples from different streams based on their common time stamps and fit all the streams with a single mixture model. Specifically, we would merge the text samples on the same time point (keeping their stream identities) to form a unified text sample on the time point. In order to match topic models across different streams, we also align the topic models from different streams; again, we keep their stream identities. Since we have kept the stream identities of both the text sample and the topic model, we can fit the right model to the right data when fitting the whole mixture model to all the streams.

We call such a mixture model a *coordinated mixture model* because the mixture models for all streams "coordinate" with each other so that each would focus more on topics that have strong correlations with topics in other streams. After we fit such a coordinated mixture model to all the streams, we will obtain all the topic models. Since the topic models from different streams are already aligned with each other in advance, we naturally obtain a correlated bursty topic pattern if all the involved topics are bursty in a similar period.

Intuitively, each topic model (i.e., word distribution) defines a soft cluster in the sense that it specifies the probability of membership of each word in the cluster, and our mixture model would group words based on their co-occurrences in samples of the same stream to form word clusters and match clusters across streams based on the correlations between the temporal distributions of the clusters from different streams.

Note that our model does not require different streams to share any vocabulary; instead it exploits the fact that topics involved in a correlated bursty topic pattern tend to have similar temporal distribution to match the topics from different streams. In Figure 1, based on the news data set used in our experiment (see Section 5), we show strong correlations of the relative frequency (i.e., frequency of a word in each time point normalized by its total frequency) distributions over time between two English words "terror" and "attack" as well as between the English word "attack" in an English news stream and its Chinese translation in a Chinese stream. In general, when the multiple streams share some common causal factors (e.g., affected by the same event), we will observe such correlations.

## 4.2 Formal Definition

We now give the formal definition of the proposed coordinated mixture model.

### 4.2.1 The Generative Model

Let $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_m\}$ be $m$ coordinated text streams with vocabularies $V_1, ..., V_m$. Without loss of generality, we assume there are $k$ correlated bursty topic patterns in our streams and associate a latent cause variable $z \in [1, k]$ with them; a different value of $z$ would indicate a different pattern. Given a value $j$ of $z$, we have a set of "aligned" topic models, each corresponding to a single stream. The topic model for stream $\mathcal{S}_i$ is given by $P(w|z = j, i)$ where $w \in V_i$. That is, the $\{P(w|z = j, i)\}_{i\in[1,m]}$ defines a potential correlated bursty topic pattern.

In order to explain *all* the words in our streams, including those that are not involved in any correlated bursty topic pattern, we further introduce a background topic model $P(w|\theta_B, i)$ for each stream $\mathcal{S}_i$.

In general, we assume that a word $w$ appearing at time $t$ in stream $\mathcal{S}_i$ with probability $P(w|t, i)$ (i.e., $w$ is a word in text sample $\mathcal{S}_{it}$) can either be a background word (thus should be generated using the background model) or potentially cover any of the $k$ patterns (thus should be generated from a mixture of the $k$ pattern models). In other words, $w$ would be regarded as a sample drawn from the following mixture model:

$$P(w|t, i) = \lambda_B P(w|\theta_B, i) + (1-\lambda_B) \sum_{z=1}^{k} P(z|t)P(w|z, i) \quad (1)$$

where $\lambda_B$ is the mixture weight of the background model, and $P(z|t)$ is the probability of choosing pattern $z$ at time point $t$.

The log-likelihood of generating text sample $\mathcal{S}_{it}$ is thus

$$\log P(\mathcal{S}_{it}) = \sum_{w\in V_i} c(w, \mathcal{S}_{it}) \log P(w|t, i)$$

where $c(w, \mathcal{S}_{it})$ is the count of word $w$ in $\mathcal{S}_{it}$. Therefore, the log-likelihood of generating all the $m$ coordinated streams is:

$$\log P(\mathcal{S}) = \sum_{t=1}^{n} \sum_{i=1}^{m} \sum_{w\in V_i} c(w, \mathcal{S}_{it}) \log P(w|t, i) \quad (2)$$

Our coordinated mixture model can be regarded as an extension of Probabilistic Latent Semantic Analysis (PLSA) [10] to model coordinated streams. Note that the $P(w|t, i)$'s for different streams are coordinated because of the common variable $t$, and $P(z|t)$, which is independent of any stream, forcing all the streams to have the same preferences for the bursty topic patterns. On the other hand, $P(w|t, i)$ is different for a different stream, allowing us to model different vocabularies.

#### 4.2.2 Parameter Estimation

We estimate the parameters of the coordinated mixture model by fitting the model to our coordinated text stream data. To model the background words in our streams (from some prior knowledge) and regularize our model, we fix our $\lambda_B$ to a constant and set

$$P(w|\theta_B, i) = \frac{c(w, \mathcal{S}_i)}{\sum_w c(w, \mathcal{S}_i)}$$

where $c(w, \mathcal{S}_i)$ is the count of word $w$ in stream $\mathcal{S}_i$.

The remaining parameters to estimate are $P(w|z, i)$ and $P(z|t)$. Without assuming any prior knowledge, we may use the maximum likelihood estimator and use the expectation-maximization (EM) algorithm to compute an estimate iteratively. The expectation step is to calculate:

$$P(z|t, w, i) = \frac{(1-\lambda_B)P^{(l)}(z|t)P^{(l)}(w|z,i)}{\lambda_B P(w|\theta_B, i) + (1-\lambda_B)\sum_z P^{(l)}(z|t)P^{(l)}(w|z,i)}.$$

The maximization step is to update the probabilities:

$$\begin{aligned} P^{(l+1)}(z|t) &= \frac{\sum_i \sum_{w\in V_i} c(w,\mathcal{S}_{it})P(z|t,w,i)}{\sum_z \sum_i \sum_{w\in V_i} c(w,\mathcal{S}_{it})P(z|t,w,i)} \\ P^{(l+1)}(w|z,i) &= \frac{\sum_t c(w,\mathcal{S}_{it})P(z|t,w,i)}{\sum_{w\in V_i}\sum_t c(w,\mathcal{S}_{it})P(z|t,w,i)} \end{aligned} \quad (3)$$

### 4.3 Constraining EM with Temporal Dependency

One deficiency of the basic coordinated mixture model is not capturing the dependency among the consecutive time points in covering topics. Specifically, each text sample makes its own, independent choice among the possible topics to cover. Intuitively, however, the text samples within a consecutive time period tend to be influenced by the same event(s). So it would be desirable to somehow force all of them to make similar choices of topics.

To implement this intuition, we propose to modify the EM algorithm to impose a temporal dependency constraint on $P(z|t)$ so that during each iteration $P(z|t)$ would be smoothed (constrained) by both its neighbors $P(z|t-1)$ and $P(z|t+1)$:

$$\begin{aligned} Q^{(l+1)}(z|t) &= \frac{\sum_i \sum_{w\in V_i} c(w,\mathcal{S}_{it})P(z|t,w,i)}{\sum_z \sum_i \sum_{w\in V_i} c(w,\mathcal{S}_{it})P(z|t,w,i)} \\ P^{(l+1)}(z|t) &= \frac{\lambda Q^{(l+1)}(z|t-1)}{2(1+\lambda)} + \frac{Q^{(l+1)}(z|t)}{1+\lambda} + \frac{\lambda Q^{(l+1)}(z|t+1)}{2(1+\lambda)} \end{aligned}$$

Here $Q^{(l+1)}(\cdot|t)$ is the original formula in EM iteration. The parameter $\lambda$ is to control the amount of smoothing. A larger $\lambda$ would impose a stronger dependency constraint among adjacent time points. When $\lambda = 0$, we impose no constraint and have $P^{(l+1)}(\cdot|t) = Q^{(l+1)}(\cdot|t)$. We found in our experiments that introducing a non-zero $\lambda$ can smooth a bursty pattern and help discover consecutive bursty periods.

### 4.4 Mutual Reinforcement across Streams

Although the discovery of correlated bursty patterns is coordinated across streams through the shared time points, the discovered topic models in each stream (i.e., $P(w|z, i)$) can be biased by some stream-specific local themes, which may peak at about the same time as the true correlated bursty topic patterns. As a result, $P(w|z, i)$ may have mixed subtopics. To "clean up" $P(w|z, i)$ and make it more focused on the correlated topics across all other streams, we propose to use a reinforcement method to reward words from different streams that are strongly correlated with each other over the entire time span.

The basic idea of this approach is to exploit the fact that those words in different streams about the same common causal factor can be expected to have strong correlations between their frequency distributions over time, whereas a noisy "local word" in a stream is unlikely to have strong correlations with them. Thus we can use the corresponding topic models in other streams (i.e., $\{P(w|z, j)\}_{j\neq i}$) to help filter out "local noise" in $P(w|z, i)$. Specifically, we would adjust $P(w|z, i)$ to promote words that are highly correlated with high probability words in all other streams. We iteratively do such adjustment for all the topic models.

To implement this idea, we first compute the global correlations between all the high probability words in one stream and those in another, where the probability of a word is given by the corresponding topic model (i.e., $P(w|z, i)$). Following [20], we represent each word from stream $\mathcal{S}_i$ by the normalized empirical frequency distribution vector over the entire time span of the stream:

$$P(t|w) = \frac{c(w, \mathcal{S}_{it})}{\sum_t c(w, \mathcal{S}_{it})}$$

We can then compute the Pearson correlation coefficient $r(x, y)$ between two words $x$ and $y$ as follows:

$$r(x,y) = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{(\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2)(\sum_{i=1}^n y_i^2 - \frac{1}{n}(\sum_{i=1}^n y_i)^2)}} \quad (4)$$

where $x_i$ and $y_i$ are $i$-th entries in $x$ and $y$'s frequency distribution vectors defined above.

With these correlation values, we then use the following iterative procedure to adjust each $P(w|z, i)$:

$$\begin{aligned} Q(w|z,i) &= P^{(l)}(w|z,i)\sum_{j:j\neq i}\sum_{w'\in V_j} r(w,w')P^{(l)}(w'|z,j) \\ P^{(l+1)}(w|z,i) &= \frac{Q(w|z,i)}{\sum_{w\in V_i} Q(w|z,i)} \end{aligned} \quad (5)$$

Intuitively, such a mutual reinforcement procedure rewards the high probability words in stream $i$ (according to $P(w|z, i)$) which have high correlations with the high probability words in its correlated topic model of other streams $P(w'|z, j)$'s.

To increase the efficiency, we can perform mutual reinforcement updating on only the words with high probabilities according to $P(w|z, i)$ and let the iteration stop after a few iterations. (In our experiments, the ranking of words according to the updated probabilities usually becomes stable after 10 iterations.)

### 4.5 Discover Correlated Busty Topic Patterns

After we get all the parameters $P(w|z, i)$ and $P(z|t)$ estimated, we can discover the correlated bursty topic patterns directly. Since $P(z|t)$ is stream independent, given a span threshold $\sigma$ and a coverage threshold $\kappa$, an identified pattern is a correlated bursty topic pattern if $P(z|t)$ satisfies the constraints in Definition 4 of bursty topic pattern.

Intuitively, $P(z|t)$ represents the strength of pattern $z$ at time $t$ and the high probability words according to $P(w|z, i)$ characterize the content of bursty pattern $z$ in stream $i$. In the experiment, we use $P(z|t)$ and $P(w|z, i)$ to represent the identified correlated bursty topic patterns. For example, in the correlated news streams in different languages where a correlated bursty topic corresponds to a real world event, $P(z|t)$'s are the intensiveness of the corresponding event over time, which can be used to identify the bursty

| | News streams | | Literature streams | |
|---|---|---|---|---|
| | English | Chinese | SIGMOD | VLDB |
| Length | 148 | 148 | 31 | 31 |
| #Word | 70,049 | 9,720 | 1,930 | 2,144 |
| #Article | 34,751 | 43,488 | 1,787 | 2,143 |
| Data size | 111MB | 63MB | 138KB | 108KB |

**Table 1: Statistics of two stream data sets**

period, and $P(w|z,i)$'s indicate what the event is about in different languages.

## 4.6 Discussion

The proposed method can work with quite large text collections. Each iteration in the EM algorithm has complexity $O(\sum_{i=1}^{m}|\mathcal{S}_i| \times k + \sum_{i=1}^{m}|V_i| \times k + n \times k)$. For the mutual reinforcement, we only need to calculate the top $N$ words of each stream given a pattern. Each iteration of reinforcement has complexity $O(N \times N \times k)$. In practice, $N = 100$ is enough since the words outside of top 100 have very low probabilities.

Although presented as a model on text streams with words as units, our model is quite general and can be applied to any discrete data streams with any interesting features as units.

# 5. EXPERIMENTS

To evaluate our methods of identifying major correlated bursty topic patterns, we conduct experiments on a news data set and a literature data set. The correlated bursty topic patterns in news streams are highly related to major real world events and the patterns in literature streams can indicate research paradigm shifts. In this section, we show that our proposed methods can identify meaningful correlated bursty topic patterns from these two types of coordinated streams to reveal the major real world events and research paradigm shifts with appropriate time line.

## 5.1 Data Sets

### 5.1.1 News Streams

Our news streams consist of six months' news articles of Xinhua English and Chinese newswires dated from June 8th, 2001 through November 7th, 2001. There are altogether 43,488 documents in Chinese and 34,751 documents in English distributed in the 148 days. Each day is used as a time point in these two streams thus the length of the coordinated streams is 148.

In each stream, a text sample with time stamp $i$ is the concatenation of all news articles appearing on date $i$. For Chinese news articles, there is no "space" between two Chinese words and the meaning of each single Chinese character can be interpreted quite differently depending on its context. The ambiguity of Chinese characters can be alleviated much by grouping them into bi-grams. Thus, without using sophisticated Chinese segmentation tools, we use bigrams of Chinese characters as words in the Chinese news stream. Therefore, at each time point (i.e., day), each news stream has a sequence of the aforementioned words which appear in the corresponding stream at that time point. As we will show, even with such crude segmentation, the mining results are already quite interesting. Naturally, with a better segmentation tool, our mining results can be further improved.

### 5.1.2 Literature Streams

The data set in literature domain we use is similar to the one used in [12]. Specifically, we use all the paper titles of SIGMOD and VLDB from the years 1975–2005 in our experiment. SIGMOD and VDLB are two major conferences in database research community, which have existed for more than 30 years since 1975. Publications accumulated in the two conferences over years naturally form coordinated text streams. In this data set, each year is treated as a time point and thus the length of the coordinated streams is 31.

The statistics of the news and literature data sets are shown in Table 1. Besides in different genres, the two data sets are clearly different in statistic measures. We use these two data sets to test the generality of our proposed methods.

## 5.2 Parameter Setting

In the coordinated mixture models, there are several user-input parameters which provide flexibility for bursty topic pattern analysis. These parameters are set empirically, as in principle, it is impossible to optimize these parameters without relying on domain knowledge. We discuss the effect of different parameters as follows.

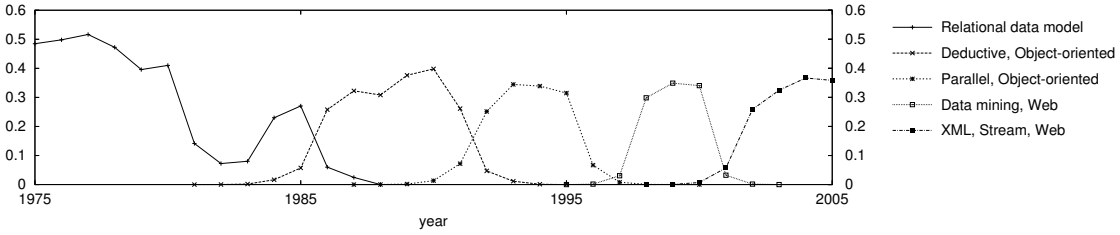
Parameter $\lambda_B$ is to control the strength of the background model. The background model captures global common words in each stream. A larger $\lambda_B$ will force discovered bursty patterns to be more discriminative from each other but an extremely large $\lambda_B$ will attract too much useful information to the background model thus weaken the interpretability of the discovered patterns. Empirically, a $\lambda_B$ suitable for text documents is between 0.9 and 0.95 and we use $\lambda_B = 0.95$ in our experiments.

Parameter $\lambda$ is to control the dependency strength among adjacent time points in the streams. A high $\lambda$ forces a high dependency, i.e., assumes that adjacent time points have very similar bursty topic patterns. A extreme lower $\lambda$ value such as 0 can not fully utilize the consecutive property of data streams. In principle, $\lambda$ can be set based on the overall similarities among adjacent time points. Empirically, we set $\lambda = 0.1$ in the following experiments.

Parameter $k$ is the number of bursty patterns shared by all coordinated streams in a data set. When no domain knowledge is available as in our experiments, we use a similar strategy as [15] to determine the number of bursty patterns by enumerating multiple possible values of $k$ and drop the patterns which do not satisfy our bursty pattern definition.

## 5.3 Results on News Data

On the news streams, we first apply our coordinated mixture model to identify the major correlated bursty topic patterns and then use our mutual reinforcement method across streams to further refine the identified patterns. Bursty patterns which satisfy $\sigma = 5$ for $\kappa = 0.01$ are kept. We finally obtain 7 major bursty topic patterns, which are shown in Figure 3. Recall that each pattern has its bursty period and a set of high probability words. We show both results in Figure 3(a) and Figure 3(b) respectively. In Figure 3(a), we plot the strength of these 7 patterns over time points indicated by probability $P(z|t)$ which is estimated from our mixture model. In Figure 3(b), we list the top 10 words with highest probabilities in each pattern mined from both En-




| Relational data model | Deductive, Object-Oriented | Parallel, Object-Oriented | Data Mining, Web | XML, Stream, Web |
|---|---|---|---|---|
| 1975-1985 | 1986-1991 | 1992-1995 | 1996-2001 | 2002-2005 |
| **design** 0.0418 | **object** 0.0486 | **object** 0.0687 | **mine** 0.0440 | **xml** 0.0772 |
| **relational** 0.0410 | **deductive** 0.0308 | **parallel** 0.0334 | **web** 0.0279 | **stream** 0.0633 |
| language 0.0333 | **orient** 0.0303 | **orient** 0.0308 | **warehouse** 0.0235 | **service** 0.0308 |
| **model** 0.0287 | **knowledge** 0.0255 | persistent 0.0177 | **dimension** 0.0234 | **web** 0.0228 |
| file 0.0274 | extend 0.0240 | multidatabase 0.0153 | similar 0.0148 | sql 0.0178 |
| base 0.0880 | **deductive** 0.0256 | **object** 0.0759 | **warehouse** 0.0366 | **xml** 0.0804 |
| **design** 0.0544 | **object** 0.0236 | **orient** 0.0453 | **mine** 0.0223 | **stream** 0.0594 |
| data 0.0375 | **orient** 0.0225 | **parallel** 0.0441 | **web** 0.0201 | **web** 0.0292 |
| **model** 0.0350 | **knowledge** 0.0204 | rule 0.0270 | **dimension** 0.0175 | **service** 0.0211 |
| **relational** 0.0341 | skew 0.0201 | composite 0.0191 | use 0.0151 | search 0.0199 |

**Figure 2: Research paradigm shifts in the literature streams. The words in upper and lower parts of the table above are from SIGMOD and VLDB streams respectively. Shared words are bold-faced.**

glish and Chinese news streams. For each Chinese bi-gram, we include its English translation in the parenthesis after itself. Note that Chinese bi-grams are not always complete Chinese words. We use "*" to indicate that a bi-gram is only a fragment of our translation.

Figure 3 gives us a good, unified summary of the two streams by the 7 identified bursty topic patterns. The 1st pattern is about Communist Party of China (CPC) and has a sharp peak around July 1st, 2001, which is the 80th anniversary of CPC. The 2nd pattern is on the bid of Olympic 2008 and bursts from July 13th, 2001 when Beijing, the Capital of China, won the bid of Olympic 2008. The 3rd pattern is the 9th swimming championship held in Fukuoka in 2001. The 4th pattern is the shrine event in Japan and the 5th pattern is the 21st Universiade held in Beijing during Aug. 22nd to Sep. 1st, 2001. 9–11 event is the 6th pattern in the figure. From Figure 3(a), we can see this pattern accurately bursts at September 11th, 2001. The set of high probability words such as "terror" and "attack" are very informative in both English and Chinese. The last pattern identified by our algorithm is Afghanistan war which happened consequently after the 9–11 event. Though both 9–11 and Afghanistan war are related to terrorists, they are two different events and our methods are able to distinguish them correctly.

The results above show that our proposed methods can successfully identify correlated bursty topic patterns in coordinated news streams. It could have many applications. For example, most of the identified English and Chinese words, although in completely different vocabularies, match very well. This can potentially help cross-lingual information retrieval and integration [22, 20]. The bursty period of a pattern can be used to analyze the trend of the cause event. Furthermore, combining the identified bursty words and periods together is informative enough to give a unified summary of the coordinated text streams.

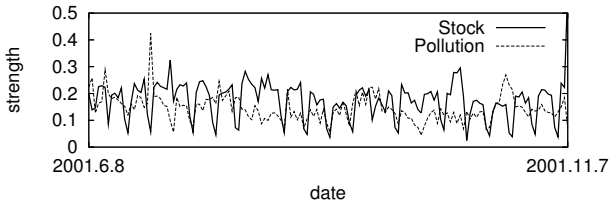



**Figure 4: Result of document-based clustering**

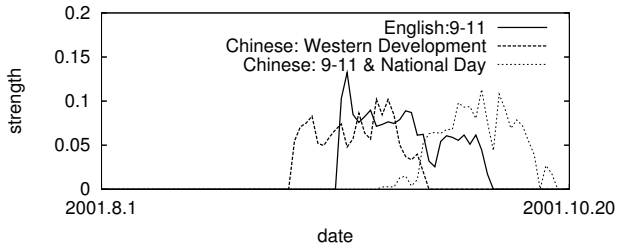
## 5.4 Results on Literature Data Set

We apply our methods to the SIGMOD and VLDB stream data, and the major bursty patterns which satisfy $\sigma = 2$ for $\kappa = 0.01$ are shown in Figure 2. The identified patterns summarize the research paradigm shifts interestingly: at the beginning, the database community focused on relational data model (1st pattern: 1975-1985); then the research focus changed to deductive and object-oriented databases (2nd pattern: 1986-1991); after that, many researchers began to study parallel databases while the object-oriented database research stayed as a major topic (3rd pattern: 1992-1995); data mining and web related topics became popular after 1996 (4th pattern: 1996-2001); in recent years, XML and stream data management became dominant together with Web related topics (5th pattern: 2002-2005).

In a whole, we can see that all the identified paradigm shifts can reflect well the real progress in the database community. This result can not only provide a big picture of the database field to a novice, but also assist an expert in writing overviews about the whole field.

## 5.5 Detailed Analysis

In this section, based on the news streams, we compare our proposed methods to directly applying PLSA-based doc-




| English | Chinese | |
|---|---|---|
| 9–11 | Western development | 9-11 & national day |
| september | 西部(the western) | 恐怖(terror) |
| attack | 发展(development) | 旅游(tourism) |
| us | 华商(Chinese merchants) | 怖主(*terrorism) |
| terrorist | 投资(investment) | 国庆(national day) |
| terror | 开发(exploiture) | 游客(tourist) |
| unite | 教育(education) | 作风(style) |
| york | 文化(culture) | 黄金(golden) |
| washington | 我国(our country) | 袭击(attack) |
| state | 基金(fund) | 月饼(moon cake) |
| new | 教师(teacher) | 金周(*golden week) |

**Figure 5: Aligning patterns across streams.**

| English | Chinese |
|---|---|
| afghanistan | 富汗(*Afghanistan) |
| taliban | 阿富(*Afghanistan) |
| **us** | 恐怖(terror) |
| terror | **会议(meeting)** |
| **anthrax** | 军事(military) |
| military | **亚太(Asian Pacific)** |
| **apec** | **作风(style)** |
| **attack** | **美国(U.S.A)** |
| strike | **组织(organization)** |
| **economic** | 打击(strike) |
| English | Chinese |
| afghanistan | 阿富(*Afghanistan) |
| taliban | 富汗(*Afghanistan) |
| **islamabad** | **对阿(*to Afghanistan)** |
| **kabul** | 军事(military) |
| **led** | **事打(*military strike)** |
| military | **事行(*military action)** |
| strike | 恐怖(terror) |
| **kandahar** | 打击(strike) |
| **civilian** | **汗的(*Afghani)** |
| terror | 喀布(Kabul) |

**Figure 6: Topic models before (top-half) and after (bottom-half) mutual reinforcement. Distinct words are bold-faced.**

ument clustering method [10, 25] and analyze different variants of our methods.

### 5.5.1 Direct application of PLSA

To discover bursty patterns, a straightforward method would be to apply document based clustering methods which have been widely used to discover topics in a document collection. However, as will be shown, these methods are not effective in detecting "bursty" patterns. In this experiment, we use the English stream and treat each English article as a document. We use PLSA based clustering method proposed in [10, 25] to group articles into clusters. The strength of a cluster in each day is calculated as in [15] and is proportional to the number of documents in the cluster which appears on that day. In Figure 4, we show the strength of two biggest clusters identified by PLSA. The first cluster is about "stock" and it has high probability words such as "stock", "dollar", and "millions". The second cluster is about the pollution in China and it has high probability words such as "pollute", "water", and "protect". From their strength over time plotted in Figure 4, we can see that neither cluster is a meaningful bursty pattern as they occur rather evenly in the whole time period. This shows that document-based clustering methods are biased by these "common" topics and are ineffective for detecting bursty patterns. One reason why this is so is because such a method does not utilize the time information in the stream. In contrast, our method utilizes time information and can detect bursty patterns as shown in Figure 3.

### 5.5.2 Coordinated streams v.s. single streams

Another possible method for finding correlated bursty topic patterns is to first find busty patterns in each single stream and then align patterns across streams according to their temporal overlaps. However, this ad-hoc method does not work well and we show an example in Figure 5. In this figure, one bursty pattern from English news stream and two bursty patterns from the Chinese news stream are shown. The pattern in English stream is the 9–11 event. The first one in Chinese stream is about "western development" and the second is a mix of 9–11 and Chinese "national day holiday" (October 1st) events. From Figure 5, we can see that the bursty periods of both Chinese patterns overlap with that of the 9-11 pattern from English news. Thus both of the Chinese patterns can be aligned with the English 9–11 pattern but neither of them is a good match. This example shows that patterns identified from single streams can be biased by their local patterns and the alignment of such biased patterns across streams could be imprecise and misleading.

Since correlated bursty patterns presumably share similar periods, the coordinated mixture model considers all the streams simultaneously and calculates the same bursty periods for their correlated bursty patterns. As shown in Figure 3(a), correlated bursty patterns have the same bursty period thus we do not need to align them. By considering coordinated stream simultaneously, our method is more robust for identifying important correlated patterns.

### 5.5.3 Effect of mutual reinforcement

We demonstrate the advantage of mutual reinforcement in Figure 6. Figure 6 (a) and Figure 6 (b) show the words of the identified patterns before and after mutual reinforcement, where the differences are bold-faced. In Figure 6 (a), the words of Afghanistan war are the majority but still mixed with some noisy words such as "APEC" (The Asia-Pacific Economic Cooperation) and "economic." With mutual reinforcement across Chinese and English streams, those noisy words are suppressed and the pattern becomes more coherent. This is because those major words in both streams are not only the high probability words, but also have high *global* temporal correlations with their counterparts in the other stream. Our mutual reinforcement procedure can effectively utilize these properties and thus improve the quality of the bursty patterns.

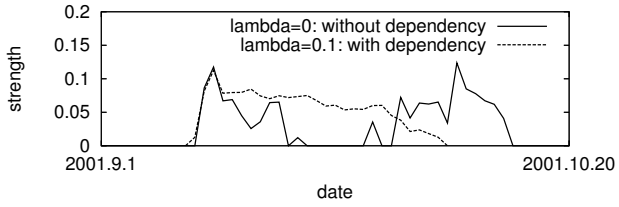



**Figure 7: Effect of local dependency.**

### 5.5.4 *Effect of local dependency*

The advantage of the constraining EM by local dependency is that it can utilize the consecutive property of a stream when parameter $\lambda > 0$. To see the effect of applying local dependency, we compare $\lambda = 0.1$ and $\lambda = 0$ in Figure 7 using the example of 9–11. It is clear that the periods are inconsecutive and rugged for $\lambda = 0$, making it hard to interpret. On the contrary, using dependency gives us consecutive bursty periods which are natural in reality.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we defined and studied a novel problem of mining correlated bursty patterns from coordinated text streams. We proposed coordinated mixture models which can identify bursty words and their bursty periods simultaneously. Furthermore, we enhanced our model by incorporating local dependency along the time line and proposed a mutual reinforcement approach across streams to further refine the correlated bursty patterns. We evaluated our methods on two data sets: coordinated news streams and coordinated literature streams. On the news streams, the identified bursty patterns reflect the real world events and on the literature streams, the identified bursty patterns well summarize the research paradigm shifts in the database community.

Besides the correlated patterns, there are local bursty patterns in each stream. In the future, we will further study the problem of identifying both global correlated and local patterns. We also plan to apply our methods to other coordinated non-text streams to study their correlated bursty patterns.

## 7. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable comments. This work is in part supported by a Microsoft Live Labs Research Grant, a Google Research Grant, and an NSF CAREER grant IIS-0347933.

## 8. REFERENCES

[1] C. Aggarwal. *Data Streams: Models and Algorithms.* Springer, 2007.

[2] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. On demand classification of data streams. In *KDD*, pages 503–508, 2004.

[3] R. Agrawal, K.-I. Lin, H. S. Sawhney, and K. Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In *VLDB*, pages 490–501, 1995.

[4] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[5] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *SIGIR*, pages 37–45, 1998.

[6] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 2003.

[7] S. Chien and N. Immorlica. Semantic similarity between search engine queries using temporal correlation. In *WWW*, pages 2–11, 2005.

[8] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In *VLDB*, pages 181–192, 2005.

[9] J. Han and M. Kamber. *Data Mining: Concepts and Techniques, 2nd Ed.* Morgan Kaufmann, 2006.

[10] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[11] G. Hulten, L. Spencer, and P. Domingos. Mining time-changing data streams. In *KDD*, pages 97–106, 2001.

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

[13] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW*, pages 568–576, 2003.

[14] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*, pages 533–542, 2006.

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

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

[17] R. Sproat, T. Tao, and C. Zhai. Named entity transliteration with comparable corpora. In *ACL*, 2006.

[18] R. Swan and J. Allan. Extracting significant time varying features from text. In *CIKM*, pages 38–45, 1999.

[19] R. Swan and J. Allan. Automatic generation of overview timelines. In *SIGIR*, pages 49–56, 2000.

[20] T. Tao and C. Zhai. Mining comparable bilingual text corpora for cross-language information integration. In *KDD*, pages 691–696, 2005.

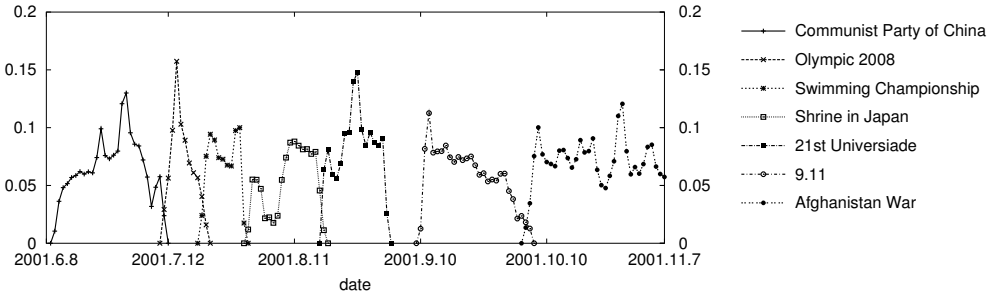
[21] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *SIGMOD*, pages 131–142, 2004.

[22] J. Xu, R. Weischedel, and C. Nguyen. Evaluating a probabilistic model for cross-lingual information retrieval. In *SIGIR*, pages 105–110, 2001.

[23] Y. Yang, T. Ault, T. Pierce, and C. W. Lattimer. Improving text categorization methods for event tracking. In *SIGIR*, pages 65–72, 2000.

[24] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *SIGIR*, pages 28–36, 1998.

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




(a) Bursty Periods

| Events | | |
|---|---|---|
| CPC(6.10-7.11) | Olympic 2008(7.11-7.21) | Swimming Championship(7.20-7.30) |
| wimbledon 0.3325 | olympic 0.4671 | fukuoka 0.4439 |
| 80th 0.2641 | 2008 0.2655 | swim 0.2349 |
| cpc 0.1983 | bid 0.1527 | 9th 0.1134 |
| communist 0.1176 | agra 0.0330 | fina 0.0812 |
| milosevic 0.0320 | congratulate 0.0205 | championship 0.0681 |
| anniversary 0.0140 | success 0.0106 | genoa 0.0152 |
| hague 0.0070 | musharraf 0.0100 | g8 0.0106 |
| leadership 0.0065 | beije 0.0099 | wahid 0.0077 |
| party 0.0060 | host 0.0092 | mpr 0.0037 |
| aids 0.0043 | toronto 0.0070 | abdurrahman 0.0034 |
| 共产(*communism) 0.2554 | 奥运(Olympics) 0.2211 | 锦赛(*championship) 0.3292 |
| 产党(*communist party) 0.2534 | 北京(Beijing) 0.1408 | 世锦(world championship) 0.2197 |
| 国共(*country communism) 0.1648 | 奥委(Olympic committee) 0.0772 | 游泳(swimming) 0.1592 |
| 党员(party member) 0.0559 | 际奥(*international Olympics) 0.0771 | 泳锦(*swimming championship) 0.0630 |
| 党的(party's) 0.0344 | 年奥(*Olympic year) 0.0685 | 泳世(*swimming championship) 0.0479 |
| 建党(party establishment) 0.0279 | 申办(bid) 0.0518 | 界游(*swimming championship) 0.0431 |
| 周年(anniversary) 0.0261 | 委会(*committee) 0.0356 | 王妍(Wang Yan) 0.0226 |
| 党成(*party achievement) 0.0236 | 成功(success) 0.0316 | 者马(*reporter Ma) 0.0142 |
| 支部(branch) 0.0175 | 京申(*Beijing's bid) 0.0299 | 由泳(*free style) 0.0136 |
| 的党(*'s party) 0.0140 | 办奥(*bid) 0.0273 | 米自(*free style) 0.0122 |

| Events | | | |
|---|---|---|---|
| Shrine in Japan(7.31-8.18) | 21st Universiade(8.18-9.2) | $9 \cdot 11(9.11 - 10.6)$ | Afghanistan War(10.5-11.7) |
| shrine 0.4813 | universiade 0.1902 | terror 0.3946 | afghanistan 0.6177 |
| yasukuni 0.3742 | game 0.1534 | terrorist 0.2280 | taliban 0.1597 |
| sadc 0.0933 | men 0.0826 | attack 0.1746 | islamabad 0.0432 |
| koizumi 0.0139 | women 0.0719 | us 0.0384 | kabul 0.0356 |
| custom 0.0132 | gold 0.0562 | laden 0.0237 | aip 0.0195 |
| jerusalem 0.0056 | university 0.0543 | bin 0.0213 | led 0.0188 |
| malawi 0.0052 | competition 0.0530 | osama 0.0166 | military 0.0187 |
| japan 0.0052 | 21st 0.0454 | washington 0.0159 | strike 0.0146 |
| orient 0.0023 | final 0.0433 | unite 0.0132 | kandahar 0.0132 |
| palestinian 0.0014 | medal 0.0306 | against 0.0098 | civilian 0.0071 |
| 参拜(visit) 0.1950 | 大运(universiade) 0.4159 | 袭击(attack) 0.3691 | 阿富(*Afghanistan) 0.2633 |
| 神社(*shrine) 0.1934 | 金牌(gold medal) 0.1253 | 恐怖(terror) 0.2995 | 富汗(*Afghanistan) 0.2612 |
| 靖国(*shrine) 0.1921 | 运动(sports) 0.0505 | 击事(*attack event) 0.1130 | 对阿(*to Afghanistan) 0.1028 |
| 国神(*shrine) 0.1911 | 国队(*national team) 0.0489 | 怖袭(*terroristic attack) 0.0643 | 军事(military) 0.0637 |
| 拜靖(*visit shrine) 0.1012 | 届大(*21st universiade) 0.0403 | 事件(event) 0.0613 | 事打(*military strike) 0.0517 |
| 日本(Japan) 0.0566 | 女子(woman) 0.0364 | 怖主(*terrorism) 0.0426 | 事行(*military action) 0.0494 |
| 围棋(I-go) 0.0388 | 生运(*universiade) 0.0309 | 纽约(New York) 0.0119 | 恐怖(terror) 0.0478 |
| 首相(Prime Minister) 0.0058 | 选手(athlete) 0.0280 | 怖分(*terroist) 0.0100 | 打击(strike) 0.0394 |
| 侵略(invasion) 0.0057 | 动会(*sports game) 0.0273 | 打击(strike) 0.0066 | 汗的(*Afghani) 0.0379 |
| 棋手(I-go player) 0.0032 | 运会(*universiade) 0.0260 | 击恐(*beat terrorism) 0.0033 | 的军('s army) 0.0251 |

(b) Bursty Words

**Figure 3: Seven events detected from the coordinated news streams. Bursty periods are in parenthesis after the labels. "*" denotes that the Chinese bi-gram is part of its translation. The upper part is English words and the lower part is the Chinese bi-grams. The number after each word is the probability $P(w|z,i)$ discussed in Section 4.5.**