A Sentence Level Probabilistic Model for Evolutionary Theme Pattern Mining from News Corpora

Shizhu Liu, Yuval Merhav, Wai Gen Yee, Nazli Goharian, Ophir Frieder
Information Retrieval Lab
Department of Computer Science
Illinois Institute of Technology
Chicago, IL, USA
{slu28, yee}@iit.edu, {yuval, nazli, ophir}@ir.iit.edu

ABSTRACT
Some recent topic model-based methods have been proposed to discover and summarize the evolutionary patterns of themes in temporal text collections. However, the theme patterns extracted by these methods are hard to interpret and evaluate. To produce a more descriptive representation of the theme pattern, we not only give new representations of sentences and themes with named entities, but we also propose a sentence-level probabilistic model based on the new representation pattern. Compared with other topic model methods, our approach not only gets each topic’s distribution per term, but also generates candidate summary sentences of the themes as well. Consequently, the results are easier to understand and can be evaluated using the top sentences produced by our probabilistic model. Experimentation with the proposed methods on the Tsunami dataset shows that the proposed methods are useful in the discovery of evolutionary theme patterns.

Categories and Subject Descriptors
H.3.3 [Knowledge Management]: Text Mining,
H.3.3 [Knowledge Management]: Semantic Techniques

General Terms
Algorithms, Experimentation.

Keywords
Temporal Text Mining, Evolutionary Theme Patterns, Topic Model, Named Entities

1. INTRODUCTION
In many application domains, we need to process document collections collected over time, in which each document bears a meaningful timestamp, such as the publication dates of news articles and scientific literature. For individuals who are charged with monitoring or reviewing a particular field, tracking changes of topics in such temporal text collections is particularly important. For example, to track an emerging event in a given news stream, it is helpful to reveal the involved persons, organizations and places of all time periods and show how they impact each other. But a manual review of thousands of news reports is simply not feasible. A system that can discover and summarize the latent temporal theme patterns and demonstrate how they evolve over time automatically would be very useful.

To implement such a system, two steps must be taken: (1) outlining the underlying thematic structures of the given collection; (2) identifying the transition process of relevant theme patterns along the timeline. Recent research work fulfill these two steps by applying topics models in temporal text collections to discover underlying thematic structures and track how latent themes evolves by comparing the discovered structures of different time intervals[6][9][10]. However, some basic drawbacks of current topic model methods hinder their incorporation into real applications: First, existing topic model methods use the top terms to represent the extracted theme. Although top terms are often intuitively meaningful, unless the user knows the source collection, it is difficult to interpret the discovered topic only by top terms. Second, the experimental results of all the existing topic model methods cannot be evaluated directly or be compared with each other.

To provide a more descriptive way to represent the theme pattern, in this paper, we propose a sentence-level probabilistic model. First, instead of representing the theme patterns only by terms, we represent sentences and themes with the assistance of named entities. Integrated names of persons, organizations, locations etc will convey more accurate and specific information to the user. Second, we assume that each sentence also reflects the mixture of underlying topics with different proportions. Thus, by estimating the parameters of our model, we can get not only each topic’s distribution per words, but the probability that each sentence belongs to each topic simultaneously. Top sentences of each theme can be used as the summary for theme at the given time interval, which greatly facilitates the user ability to understand the extracted theme. Moreover, evaluation of the extracted theme patterns becomes feasible by judging the relevance of those top sentences.

The remainder of this paper is organized as follows. We review previous research in Section 2. In Section 3, we study the characteristics of news articles and represent their sentences with named entities. Based on this representation, we propose the sentence-level probabilistic model. In Section 4, we introduce our strategy of identifying the theme thread over time. In Section 5, we describe our experimental methodology and report our results. We conclude our paper and discuss the future work in Section 6.
2. Related Work

Some research works have been done to mine useful temporal information from temporal text collections. However, most of them are performed on document level.

Emerging trend detection is a related subproblem of temporal text mining. Existing works focus on the detection of novel topics and trends in text streams and discovering temporal structures in the streams rather than summarizing the evolutionary process of the themes [11, 12, 13, 14, 15, 16, 18, 19, 20]. Topic detection and tracking (TDT), a research program that investigates methods for automatically organizing news stories by the events that they discuss [22, 23, 24, 25], is another related work. Two closely related subproblems of TDT are called topic tracking and new event detection, respectively. The goal of topic tracking is to classify news reports into one of several topics specified by a user, while the goal of new event detection (formerly called first story detection) is to identify texts that discuss a topic that has not already been reported in earlier texts.

In the field of text clustering, the topic models studied [2, 3, 5] are similar to the mixture theme model we use here. Their goal is to discover underlying sub-clusters of documents. Like most existing clustering methods, these works do not consider the variations of subtopics in different time periods. In the field of text summarization, Allan et al. proposed several methods for constructing temporal summaries of news stories based on ranking and selecting sentences obeying a temporal order [1]. The purpose of the summarization work is to sketch the most important points of the text stream.

Unlike the above mentioned work, our task aims at mining the latent themes patterns from given text streams and revealing how they evolve over time. Recent work based on the “topic model” makes it possible to track the variations of subtopics in different time periods at the word level. Most of these works use top words to represent extracted subtopics. A major difference of existing models is their time model.

Some research work relies on Markov assumptions of discretization of time [17]. For instance, Dynamic Topic Models (DTMs) presented by Blei and Lafferty capture the alignment among topics across time steps by using a Kalman filter [6][7]. Mei and Zhai present a method based on Hidden Markov Models (HMMs) [9] to model and decode the shift between trans-collection themes in the whole collection.

Wang and McCallum present Topics over Time (TOT) [10], a non-Markov continuous time model of topical trends. Their work assumes that the meaning of a particular topic can be relied upon as constant, but the topics’ occurrence and correlations change significantly over time. Their model parameterizes a continuous distribution over time associated with each topic, and topics are responsible for generating both observed timestamps as well as words.

3. Sentence Level Probabilistic Model

Our system processes the temporal collection with the following three steps: 1) Divide the collection into a sequence of slices with a given granularity; 2) Mine the theme pattern and summary sentences in each time interval; 3) Identify the evolutionary transition of theme spans over time. Step 1 is trivial and Step 2 is based on the sentence level probabilistic model introduced in Section 3. The method of discovering evolutionary transitions over the extracted theme spans of all the time intervals will be introduced in Section 4.

3.1 Representations of Sentence and Theme

A sentence is a basic semantic unit. Compared with single words, each sentence represents an integrated semantic idea. Compared with entire documents, each sentence is apt to serve one topic one time. Thus, a document consists of a group of topics, each of which can be represented by a group of independent sentences.

Inspired by existing work on news corpora that make use of categorized entities to represent news articles [13], our model represents a sentence as a mixture of three kinds of information: who (persons or organization), where (locations) and what (keywords). For each sentence, the keywords represent the remaining contents after removing named entities and stop words. Usually, there are many named entities and keywords in news articles, and we generally refer to them as “entities” in this paper.

Themes can be represented in a similar way.

\[
\text{sentence} = \{\text{persons}, \text{locations}, \text{keywords}\} \\
\text{theme} = \{\text{persons}, \text{locations}, \text{keywords}\}
\]

Here, sentences and themes of the given collections will be represented by three kinds of entities, and each kind of entity has its own term space. For simplicity, we follow the assumption presented in [13], that the named entities in a sentence are independent.

\[
p(\text{sentence}) = p(\text{persons})p(\text{locations})p(\text{keywords})
\]

3.2 The Generative Model

Based on the representation of sentences and themes introduced in Section 3.1, we model the generation process of the sentence by a graph model. Supposing that there exists a given number of themes shared by all the sentences in the collection, the three categories of entities of each sentence are assumed to be independently drawn from a mixture of multinomials. The mixing of each sentence are randomly drawn. Thus, for the \(j\)th theme \(z_j\), the two-step generating process of a news sentence \(s_i\) is:

1. Choose a theme \(z_j \sim \text{Multinomial}(\theta_j)\)
2. Generate a sentence \(s_i \sim P(s_i|z_j)\). For each entity of it, according to the type of current entity:
   a. Choose a person \(\sim \text{Multinomial}(\theta_{p,i})\)
   b. Choose a location \(\sim \text{Multinomial}(\theta_{l,i})\)
   c. Choose a keyword \(\sim \text{Multinomial}(\theta_{k,i})\)

See, I don’t know what the difference between “temporal structures” and “summarizing the evolutionary process” is. Hopefully, someone in the area does know.

So, what is proposed in this paper is “summarizing the evolutionary process”?2
The generative process is shown in Figure 1. The graphical model for this generative process is shown in Figure 1. $N$ is the term space size of the three kinds of entities and $M$ is the number of the sentences in the collections.

The Maximum Likelihood method is used here to estimate the parameters in the model, by introducing a latent variable, topics, we can give the log-likelihood of the joint distribution as:

$$
\log(\prod_{i=1}^{M} p(z_i | \theta)) = \sum_{i=1}^{M} \log(\prod_{j=1}^{T} p(s_j | z_i, \theta))
$$

where $S$ represents the corpus of sentences in the documents; $M$ is the number of sentences in the news corpus and $T$ is the number of topics.

Given a topic $j$, the three kinds of information of the $i$\textsuperscript{th} sentence are conditionally independent:

$$
P(s_i | z_j) = P(\text{locations} | z_j) P(\text{persons} | z_j) P(\text{keywords} | z_j)
$$

Likewise, for the corresponding $i$\textsuperscript{th} document containing $X$ sentences:

$$
P(d_i | z_j) = \prod_{x=1}^{X} P(s_j | z_j)
$$

We apply the Expectation Maximization (EM) algorithm to find maximum likelihood estimates of parameters in the probabilistic models. In the E-step, we compute each sentence’s posterior probability that it belongs to the given topics, $P(z_i | x^{(i)})$, and the corresponding document’s posterior probability that it belongs to the given topics, $P(z_j | d_i)$, by:

$$
P(z_i | x^{(i)}) = \frac{P(z_i, x^{(i)})}{P(x^{(i)})} = \frac{P(z_i) P(x^{(i)} | z_i)}{P(x^{(i)})} = \frac{P(z_i) P(x^{(i)} | z_j)}{\sum_{z_j} P(z_j) P(x^{(i)} | z_j)}$$

where the superscript $(i)$ indicates the $i$\textsuperscript{th} iteration.

In the M-step, we update each topic’s distribution over all the entities and terms. For the three mixtures of unigram models, parameters are updated by:

$$
P(w_n | s_j) = \frac{1 + \sum_{i=1}^{M} P(z_i | s_j) x^{(i)} * \text{ef}(w_n)}{V + \sum_{i=1}^{M} P(z_i | s_j) x^{(i)} * \text{len}(s_j)}
$$

where $\text{ef}(w_n)$ is the count of entity $w_n$ in sentence $s_j$, len($s_j$) is the length of the sentence $s_j$, and $V$ is the vocabulary size. Laplace smoothing [8] is used here to prevent zero probabilities for infrequently occurring entities in (6).

Furthermore, we update the mixture proportions of underlying topics by:

$$
P(z_j | s_j) = \frac{\sum_{i=1}^{M} P(z_j | s_j) x^{(i)}}{M}
$$

By estimating the parameters of the sentence level probabilistic model, we not only estimate the given topic’s distributions over each entity, but also the probability that each sentence belongs to the given topics. The top entities of persons, locations and keywords can be used to represent the theme pattern, and the top sentences of each topic can be the candidate summary of the corresponding topic.

### 4. Discovering Evolutionary Theme Patterns

In our system, the evolutionary process is modeled in a context-sensitive way with a Hidden Markov Model, and the evolutionary distance between two theme spans is computed with the Kullback-Leibler divergence [4]. For any pair of theme spans $\lambda_1$ and $\lambda_2$, where $t(\lambda_1) < t(\lambda_2)$, we compute the KL-divergence as:

$$
D_{KL}(\lambda_1 \parallel \lambda_2) = \sum_{i=1}^{M} p(w_i | \lambda_1) \log \frac{p(w_i | \lambda_1)}{p(w_i | \lambda_2)}
$$

If $D_{KL}(\lambda_1 \parallel \lambda_2)$ is above a given threshold, $\lambda_2$ will be identified as evolved from $\lambda_1$. After identifying all of the evolutionary transitions over all of the time intervals, we get a set of theme threads, and each of them demonstrates the evolutionary process of its corresponding topic.

### 5. EXPERIMENTS AND RESULTS

#### 5.1 Data Set

Our experiments are carried out on the Tsunami news data set, created by Mei and Zhai [9]. This data set consists of news articles about the 2004 tsunami from Dec. 19 2004 to Feb. 8 2005. There are 7468 news articles from 10 selected sources, with the keyword query “tsunami”. In this work we chose to identify and classify three types of entities: Person, Organization, and Location.
Figure 3. A theme span extracted from Asia Tsunami data by Mei and Zhai’s method

Table 1. Performance of the Lingpipe Entity Recognizer on 15 Randomly Chosen Articles from the Tsunami Dataset

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORGANIZATION</td>
<td>0.61</td>
</tr>
<tr>
<td>LOCATION</td>
<td>0.80</td>
</tr>
<tr>
<td>PERSON</td>
<td>0.74</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>0.72</td>
</tr>
</tbody>
</table>

using LingPipe [21]. Since named entities have a substantial role in our probabilistic model, it is important that the supervised learning recognizer provided by LingPipe achieves satisfactory results. Table 1 lists both the precision and recall of the three entity types Person, Organization, and Location, manually evaluated on 15 randomly selected articles from the Tsunami news data set. As expected, our results are roughly aligned with previous works on various news data sets that report an average precision ranging from 70% to 90% and a recall of 50% [26]. An average precision of 72% here is satisfactory for our task.
5.2 Experiments on Asia Tsunami
After partitioning the collection into time intervals, we processed each news article of a given time interval according to the following steps: (1) Detect the boundaries of sentences; (2) For each sentence, extract three kinds of named entities: person, organization, location and other keywords. Use extracted entities and keywords to construct the sentence probabilistic model. After processing all of the news articles, we discovered the themes on each time intervals and identified the theme evolution threads. The results are discussed below.

5.2.1 Representations of the Evolutionary Theme Span
Figure 3 presents a theme thread extracted by Mei and Zhai’s method on the same dataset. Apparently, some of the top terms are actually parts of location names and organizations, which could be of a great confusion for the user. For example, top terms “Hong” and “Kong” listed in Jan/05/04 – Jan/22/05 are actually parts of location “Hong Kong”, top terms “Paris” and “Club” listed in Dec/28/04 – Jan/14/05 are actually parts of the organization “Paris Club”. Compared with their results, Figure 4 demonstrates a segment of a theme evolution thread extracted by our method concerning “aid from the world”. Unfortunately, space limitations prevent us from showing all of the integrated evolutionary theme threads. Three kinds of information about the theme listed here indicate people and organizations involved in the aid work, mainly countries and districts that receive the aid, and the mainly forms of aid. Furthermore, the corresponding top sentences help the user understand the story reported on the given day. Among the three kinds of entities of the given segment, entities of type person shows greater changes and the corresponding top sentences reflect such variety. For example, a new organization “Paris Club” occurred in the person space of Dec 28th 2004 and the corresponding first sentence is about the Paris Club’s plan to freeze payments to the stricken countries. With categorized entities and summary sentences, it becomes much easier for the user to understand the meaning of the theme and get to know how the story developed over time.

5.2.2 Evaluation of Theme Thread
Another advantage of the sentence level probabilistic model is that it makes the evaluation of the extracted topics feasible. In our experiments, the topics of the extracted theme threads are determined manually first. Then the extracted theme threads can be evaluated by judging the relevance of the top sentences. (i.e., we will judge if of all the top N sentences are related to the summarized topic). This would give each sentence a binary score: 1 if it should be in the topic and 0 otherwise. Accordingly, the precision for the top N sentences will be computed for the extracted theme threads. We choose to evaluate the discovered theme by judging the relevance of the top sentences. For instance, the following two sentences are relevant to the topic than the 17th sentence. We compute the quality of our model by the precision at N metric.

5.2.2.1 Results Analysis
Figure 5 shows the precision for the top N sentences for three themes: world aid, research work, and personal experience. By analyzing the results, we found that topics with diverse content are apt to yield lower precision than more specific themes. For example, the precision of theme “Personal Experience” is lower than that the other two as it covered a variety of content. Personal injuries, vision of disaster, etc., can all be treated as descriptions of personal experiences in the tsunami. Thus, the terms used for personal experiences are more “general” than the terms of other two themes. As a result, some irrelevant sentences also achieve high posterior probability belongs to the theme.

Also notice that precision increases with N for the themes “Research” and “Personal Experience”. This unusual phenomenon stems from the fact that the lengths of sentences are relatively short and contain fewer words compared with the document. During parameter estimation process, the inaccurate distributions of the words will lead to wrong posterior probability of the sentence much easier.

![Figure 5. Precision at top 25 sentences of three extracted theme threads from Asia Tsunami data](image)

6. CONCLUSIONS
In this paper, we propose a sentence level probabilistic model to discover evolutionary theme patterns from temporal text collections. First, by introducing named entities, we represent sentences and themes in the news reports with three kinds of information according to the features of news articles. Entities of each sentence are assumed to be independently drawn from a mixture of multinomials, and each sentence reflects the mixture of underlying topics with different proportions. Experiments were conducted on the Tsunami data sets, which include a stream of 50 day's news articles about the tsunami disaster that happened in Asia. Extracted themes represented with named entities conveyed more specific information. Moreover, sentences with a high probability of belonging to each theme can be selected as the summary of the corresponding theme. This not only helps us understand the extracted themes, but also make it possible to evaluate the discovered theme by judging the relevance of the top sentences.
7. ACKNOWLEDGMENTS
Our thanks to all the students of IIT Information Retrieval Lab gave us constructive suggestions. The insight comments from anonymous reviewers are greatly appreciated.

8. REFERENCES