Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining

Qiaozhu Mei
Department of Computer Science
University of Illinois at Urbana Champaign

ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana Champaign

ABSTRACT

Temporal Text Mining (TTM) is concerned with discovering temporal patterns in text information collected over time. Since most text information bears some time stamps, TTM has many applications in multiple domains, such as summarizing events in news articles and revealing research trends in scientific literature. In this paper, we study a particular TTM task – discovering and summarizing the evolutionary patterns of themes in a text stream. We define this new text mining problem and present general probabilistic methods for solving this problem through (1) discovering latent themes from text; (2) constructing an evolution graph of themes; and (3) analyzing life cycles of themes. Evaluation of the proposed methods on two different domains (i.e., news articles and literature) shows that the proposed methods can discover interesting evolutionary theme patterns effectively.

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

General Terms: Algorithms

Keywords: Temporal text mining, evolutionary theme patterns, theme threads, clustering

1. INTRODUCTION

In many application domains, we encounter a stream of text, in which each text document has some meaningful time stamp. For example, a collection of news articles about a topic and research papers in a subject area can both be viewed as natural text streams with publication dates as time stamps. In such stream text data, there often exist interesting temporal patterns. For example, an event covered in news articles generally has an underlying temporal and evolutionary structure consisting of themes (i.e., subtopics) characterizing the beginning, progression, and impact of the event, among others. Similarly, in research papers, research topics may also exhibit evolutionary patterns. For example, the study of one topic in some time period may have influenced or stimulated the study of another topic after the time period. In all these cases, it would be very useful if we can discover, extract, and summarize these evolutionary theme patterns (ETP) automatically. Indeed, such patterns not only are useful by themselves, but also would facilitate organization and navigation of the information stream according to the underlying thematic structures.

Consider, for example, the Asian tsunami disaster that happened in the end of 2004. A query to Google News (http://news.google.com) returned more than 80,000 online news articles about this event within one month (Jan.17 through Feb.17, 2005). It is generally very difficult to navigate through all these news articles. For someone who has not been keeping track of the event but wants to know about this disaster , a summary of this event would be extremely useful. Ideally, the summary would include both the major subtopics about the event and any threads corresponding to the evolution of these themes. For example, the themes may include the report of the happening of the event, the statistics of victims and damage, the aids from the world, and the lessons from the tsunami. A thread can indicate when each theme starts, reaches the peak, and breaks, as well as which subsequent themes it influences. A timeline-based theme structure as shown in Figure 1 would be a very informative summary of the event, which also facilitates navigation through themes.

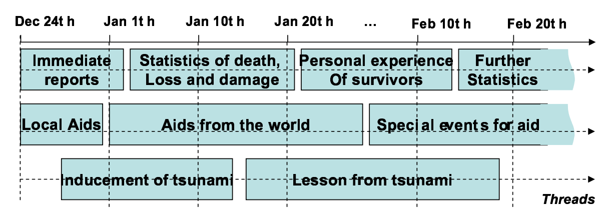

Figure 1: An example of theme thread structure

In addition to the theme structure, revealing the strength of a theme at different time periods, or the "life cycle" of a theme, is also very useful. Consider another scenario in the literature domain. There are often hundreds of papers published annually in a research area. A researcher, especially a beginning researcher, often wants to understand how the research topics in the literature have been evolving. For example, if a researcher wants to know about information retrieval, both the historical milestones and the recent research trends of information retrieval would be valuable for him/her. A plot, such as the one shown in Figure 2, which visualizes the evolution patterns of research topics, would not only serve as a good summary of the field, but also make it much easier for the researcher to selectively choose appropriate papers to read based on his/her research interests.

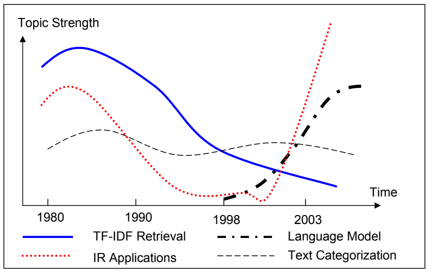

Figure 2: An example of theme strength in IR

In both scenarios, we clearly see a need for discovering evolutionary theme patterns in a text stream. In general, it is often very useful to discover the temporal patterns that may exist in a stream of text articles, a task which we refer to as *Temporal Text Mining* (TTM). Since most information bears some kinds of time stamps, TTM can be expected to have many applications in multiple domains.

Despite its importance, however, TTM has not been well addressed in the existing work. Most existing text mining work does not consider the temporal structures of text [7, 8]. There are some previous studies on TTM [10, 19, 11, 14], but the proposed methods are generally inadequate for generating the evolutionary theme patterns as shown in the two examples above. A detailed discussion of related work is given in Section 6.

In this paper, we study the problem of discovering and summarizing the ETPs in a text stream. We define this problem and present general probabilistic methods for solving the problem through (1) discovering latent themes from text, which includes both interesting global themes and salient local themes in a given time period; (2) discovering theme evolutionary relations and constructing an evolution graph of themes; and (3) modeling theme strength over time and analyzing the life cycles of themes. We evaluate the proposed methods on two data sets – a collection of 50 day's worth of news articles about the tsunami event (Dec.19, 2004 – Feb.08, 2005) and the abstracts of the ACM KDD conference papers from 1999 through 2004. The results show that our methods can discover many interesting ETPs from both data sets. In addition to news summarization and literature mining, the proposed TTM methods are also directly applicable to many other application domains, such as email analysis, mining user logs, mining customer reviews.

The rest of the paper is organized as follows. In Section 2, we formally define the general problem of ETP discovery. In Section 3, we present our approaches to extracting themes and constructing a theme evolution graph. In Section 4, we further present a hidden Markov model based method for analyzing the life cycles of themes. We discuss our experiments and results in Section 5. Finally, Section 6 and Section 7 are related work and conclusions, respectively.

2. PROBLEM FORMULATION

The general problem of ETP discovery can be formulated as follows.

Suppose we have a collection of time-indexed text documents, $C = \{d_1, d_2, ..., d_T\}$, where d_i refers to a document with time stamp i. Each document is a sequence of words from a vocabulary set $V = \{w_1, ..., w_{|V|}\}$. We define the following concepts.

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

Using a word distribution to model topics is quite common in information retrieval and text mining [5, 9, 2]. High probability words of such a distribution often suggest what the theme is about. For example, a theme about the aid from the US to help recovery from the tsunami disaster may have high probabilities for words such as "U.S.", "million", "aid", "Bush", etc.

Definition 2 (Theme Span) A ***theme span*** γ is defined as a theme θ spanning a given time interval l and is represented by $\langle\theta, s(\gamma), t(\gamma)\rangle$, where $s(\gamma)$ and $t(\gamma)$ are the starting and termination time stamps of l, respectively.

A theme span is a useful concept for associating time with themes. For the purpose of TTM, a theme is almost always tagged with a time span. We thus use "theme" and "theme span" interchangeably whenever there is no ambiguity. We call a theme span that spans the entire text stream a ***trans-collection theme***. Thus if $\gamma = \langle\theta, s, t\rangle$ is a trans-collection theme, we must have $s = 1$ and $t = T$. We use Γ to denote the set of all theme spans.

Definition 3 (Evolutionary Transition) Let $\gamma_1 = \langle\theta_1, s(\gamma_1), t(\gamma_1)\rangle$ and $\gamma_2 = \langle\theta_2, s(\gamma_2), t(\gamma_2)\rangle \in \Gamma$ be two theme spans. If $t(\gamma_1) \leq s(\gamma_2)$ (γ_1 terminates before γ_2 starts) and the similarity between theme span γ_1 and γ_2 is above a give threshold, we say that there is an ***evolutionary transition*** from γ_1 to γ_2, which we denote by $\gamma_1 \prec \gamma_2$. We also say that θ_2 is evolved from θ_1, or θ_1 evolves into θ_2. We use $\mathcal{E} \subset \Gamma \times \Gamma$ to denote all the evolutionary transitions, so that if $(\gamma_1, \gamma_2) \in \mathcal{E}$, then $\gamma_1 \prec \gamma_2$.

The concept of evolutionary transition is useful for describing the evolution relations between theme spans. With this concept, we can now define a particularly interesting theme pattern called a *theme evolution thread*.

Definition 4 (Theme Evolution Thread) Let Γ be a set of theme spans, a ***theme evolution thread*** is a sequence of theme spans $\gamma_0, \gamma_1, ..., \gamma_n \in \Gamma$ such that $(\gamma_i, \gamma_{i+1}) \in \mathcal{E}$.

Intuitively, a theme evolution thread characterizes how a family of related themes evolve over time. Since a text stream generally has multiple such theme threads, we now define another concept called *theme evolution graph* to characterize the overall theme evolution patterns of a text stream.

Definition 5 (Theme Evolution Graph) A ***Theme Evolution Graph*** is a weighted directed graph $G = (N, E)$ in which each vertex $v \in N$ is a theme span, and each edge $e \in E$ is an evolutionary transition. The weight on an edge indicates the evolution distance. Clearly, each path in a theme evolution graph represents a theme evolution thread.

An example of a theme evolution graph is shown in Figure 3, where each vertex is a theme span extracted from a subcollection obtained through non-overlapping partitioning of the stream into n sliced intervals. Each edge is an evolutionary transition. The thickness of an edge indicates how close the two themes being connected are and how trustful

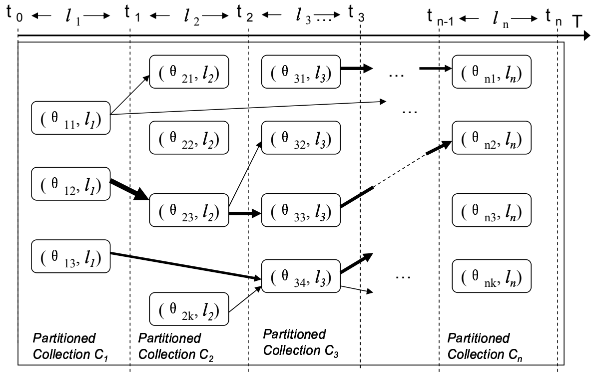

Figure 3: An example of a theme evolution graph

the corresponding evolutionary transition is; a thicker edge indicates a closer distance between the themes and a more trustful transition. For example, the distance of $\theta_{12} \prec \theta_{23}$ is smaller than that of $\theta_{11} \prec \theta_{21}$, and the former is more trustful. We also see a theme evolution thread from θ_{12}, through θ_{23}, and all the way to θ_{n2}.

Given a text stream C, a major task of the general **ETP discovery** problem is to extract a theme evolution graph from C automatically. Such a graph can immediately be used as a summary of the themes and their evolution relations in the text stream, and can also be exploited to organize the text stream in a meaningful way. Sometimes, a user may be interested in a specific theme. For example, a researcher may be interested in a particular subtopic. In this case, it is often useful to analyze the whole "life cycle" of a theme thread. Thus another task of ETP discovery is to compute the strength of a theme at different time periods so that we can see when the theme has started, when it is terminated, and whether there is any break in between.

The ETP discovery problem is challenging in many ways. First, it is a completely unsupervised task; there's no training data to discriminate theme spans. This indicates a great advantage of any techniques for ETP discovery – no/minimum prior knowledge about a domain is assumed. Second, compared with the problem of novelty detection and event tracking, which aims to segment the text and find the boundaries of events [3, 18, 13], the ETP discovery problem involves a more challenging task of modeling the multiple subtopics at any time interval for an event, and aims to discover the changing and evolutionary relations between the theme spans. Finally, the analysis of theme life cycles requires the system to decode the whole collection with themes and model the strength variations of each theme along the time line in a completely unsupervised way.

In the next two sections, we propose and present probabilistic approaches for discovering ETPs and analyzing the life cycles of themes, respectively.

3. EVOLUTION GRAPH DISCOVERY

Given a stream of text $C = \{d_1, d_2, ..., d_T\}$, our goal is to extract a theme evolution graph from C automatically. At a high-level, our methods involve the following three steps:

1. Partition the documents into n possibly overlapping subcollections with fixed or variable time intervals so that $C = C_1 \cup ... \cup C_n$ and $C_i = \{d_{t_i}, ..., d_{t_i+l_i-1}\}$ is a subcollection of l_i documents in the time span $[t_i, t_i + l_i - 1]$. In general, $t_i < t_{i+1}$, but it may be that $t_i + l_i - 1 > t_{i+1}$, since C_i's may be overlapping. The actual choice of the interval lengths l_i and whether C_i's should overlap are determined by specific applications.

2. Extract the most salient themes $\Theta_i = \{\theta_{i,1}, ..., \theta_{i,k_i}\}$ from each subcollection C_i using a probabilistic mixture model.

3. For any themes in two different subcollections, $\theta_1 \in \Theta_i$ and $\theta_2 \in \Theta_j$ where $i < j$, decide whether there is an evolutionary transition based on the similarity of θ_1 and θ_2.

Step 1 is trivial; below we describe Steps 2 & 3 in detail.

3.1 Theme Extraction

We extract themes from each subcollection C_i using a simple probabilistic mixture model as described in [20]. In this method, words are regarded as data drawn from a mixture model with component models for the theme word distributions and a background word distribution. Words in the same document share the same mixing weights. The model can be estimated using the Expectation Maximization(EM) algorithm [6] to obtain the theme word distributions.

Specifically, let $\theta_1, ..., \theta_k$ be k theme unigram language models (i.e., word distributions) and θ_B be a background model for the whole collection C. A document d is regarded as a sample of the following mixture model:
$p(w : d) = \lambda_B p(w|\theta_B) + (1 - \lambda_B) \sum_{j=1}^{k} [\pi_{d,j} p(w|\theta_j)]$
where w is a word in document d, $\pi_{d,j}$ is the mixing weight for document d for choosing the j-th theme θ_j such that $\sum_{j=1}^{k} \pi_{d,j} = 1$, and λ_B is the mixing weight for θ_B. The purpose of using a background model θ_B is to make the theme models more discriminative; since θ_B gives high probabilities to non-discriminative and non-informative words, we expect such words to be accounted for by θ_B and thus the theme models to be more discriminative. θ_B is estimated using the whole collection C as $p(w|\theta_B) = \frac{\sum_{i=1}^{T} c(w,d_i)}{\sum_{w \in V} \sum_{i=1}^{T} c(w,d_i)}$

The additional parameters to estimate are $\Lambda = \{\theta_j, \pi_{d,j} | d \in C_i, 1 \le j \le k\}$. The log-likelihood of C_i, $\log p(C_i|\Lambda)$ is

$$\sum_{d \in C_i} \sum_{w \in V} [c(w,d) \log(\lambda_B p(w|\theta_B) + (1-\lambda_B) \sum_{j=1}^{k} (\pi_{d,j} p(w|\theta_j)))]$$

where $c(w, d)$ is the count of word w in document d.

According to the EM algorithm, we can use the following iterative updating formulas to estimate all the parameters. $\{z_{d,w}\}$ is a hidden variable and $p(z_{d,w} = j)$ indicates that the word w in document d is generated using theme j given that w is not generated from the background mode.

$$p(z_{d,w} = j) = \frac{\pi_{d,j}^{(n)} p^{(n)}(w|\theta_j)}{\sum_{j'=1}^{k} \pi_{d,j'}^{(n)} p^{(n)}(w|\theta_{j'})}$$

$$p(z_{d,w} = B) = \frac{\lambda_B p(w|\theta_B)}{\lambda_B p(w|\theta_B) + (1 - \lambda_B) \sum_{j=1}^{k} \pi_{d,j}^{(n)} p^{(n)}(w|\theta_j)}$$

$$\pi_{d,j}^{(n+1)} = \frac{\sum_{w \in V} c(w,d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j)}{\sum_{j'=1}^{k} \sum_{w \in V} c(w,d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j')}$$

$$p^{(n+1)}(w|\theta_j) = \frac{\sum_{d \in C_i} c(w,d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j)}{\sum_{w' \in V} \sum_{d \in C_i} c(w',d)(1 - p(z_{d,w'} = B)) p(z_{d,w'} = j)}$$

The algorithm is only guaranteed to find a local maximum of the likelihood. We use multiple trials to improve the local maximum we obtain. We use $\frac{1}{|C_i|}\sum_{d\in C_i}\pi_{d,j}$ to measure the salience of theme j in C_i and select the most salient themes from C_i by using an empirically set threshold. We obtain the theme spans for C_i by attaching the time span of C_i to all the selected salient themes.

The same model can be applied to the whole collection C to extract trans-collection themes; we will do that in Section 4 to analyze the life cycles of trans-collection themes.

3.2 Evolutionary Transition Discovery

With the theme spans extracted from all the subcollections, we now turn to the discovery of evolutionary transitions. To discover any evolutionary transition between two theme spans, we use the Kullback-Leibler divergence [4] to measure their evolution distance. Let $\gamma_1 = \langle\theta_1, s(\gamma_1), t(\gamma_1)\rangle$ and $\gamma_2 = \langle\theta_2, s(\gamma_2), t(\gamma_2)\rangle$ be two theme spans where $t(\gamma_1) \leq s(\gamma_2)$. We assume that γ_2 has a smaller evolution distance to γ_1 if their unigram language models θ_2 and θ_1 are closer to each other. Since the KL-divergence $D(\theta_2||\theta_1)$ can model the additional new information in θ_2 as compared to θ_1, it appears to be a natural measure of evolution distance between two themes.

$$D(\theta_2||\theta_1) = \sum_{i=1}^{|V|} p(w_i|\theta_2)\log\frac{p(w_i|\theta_2)}{p(w_i|\theta_1)}$$

Note that the KL-divergence is asymmetric and it makes more sense to use $D(\theta_2||\theta_1)$ than $D(\theta_1||\theta_2)$ to measure the evolution distance from θ_1 to θ_2.

For every pair of theme spans γ_1 and γ_2 where $t(\gamma_1) \leq s(\gamma_2)$, we compute $D(\theta_2||\theta_1)$. If $D(\theta_2||\theta_1)$ is above a threshold ξ, we will infer that $\gamma_1 \prec \gamma_2$. The threshold ξ allows a user to flexibly control the strength of the theme transitions.

Once we extract the theme spans from all the subcollections and identify all the evolutionary transitions, we essentially have a theme evolution graph.

4. ANALYSIS OF THEME LIFE CYCLES

The theme evolution graph discussed above gives us a microcosmic view of the ETPs – revealing the major theme spans within each time interval and their evolutionary structures. To obtain a macroscopic view of the ETPs, it would be useful to extract the global evolutionary patterns of themes over the whole text stream and analyze the "life cycle" of each specific theme.

Definition 6 (Theme Life Cycle) Given a text collection tagged with time stamps and a set of trans-collection themes, we define the ***Theme Life Cycle*** of each theme as the strength distribution of the theme over the entire time line. The strength of a theme at each time period is measured by the number of words generated by this theme in the documents corresponding to this time period, normalized by either the number of time points (giving an *absolute strength*), or the total number of words in the period (giving a *relative strength*). The absolute strength measures the absolute amount of text which a theme can explain, while the relative strength indicates which theme is relatively stronger in a time period.

We now present a method based on Hidden Markov Models (HMMs) [17] to model and decode the shift between trans-collection themes in the whole collection. Based on the decoding results, we can then compute the theme strengths and analyze theme life cycles in a straightforward way.

We first give a brief introduction to HMMs. An HMM can be characterized by a set of hidden states $S = \{s_1, ..., s_n\}$, a set of observable output symbols $O = \{o_1, ..., o_m\}$, an initial state probability distribution $\{\pi_i\}_{i=1}^n$, a state transition probability distribution $\{a_{i,j}\}_{j=1}^n$ for each state s_i, and an output probability distribution $\{b_{i,k}\}_{k=1}^m$ for each state s_i. An HMM defines a generative probabilistic model for any sequence of symbols from O with parameters satisfying the following constraints: (1) $\sum_{i=1}^n \pi_i = 1$; (2) $\sum_{j=1}^n a_{i,j} = 1$; (3) $\sum_{k=1}^m b_{i,k} = 1$.

To model the theme shifts in our text stream, we assume that the collection, which is represented as a long sequence of words, is stochastically generated from an HMM constructed in the following way. We first extract k trans-collection themes from the collection using the mixture model described in the previous section. We then construct a fully connected HMM with $k + 1$ states, of which k states correspond to the extracted k themes and the other one corresponds to a background theme language model estimated based on the whole collection. The entire vocabulary V is taken as the output symbol set, and the output probability distribution of each state is set to the multinomial distribution of words given by the corresponding theme language model. A 3-theme HMM is shown in Figure 4.

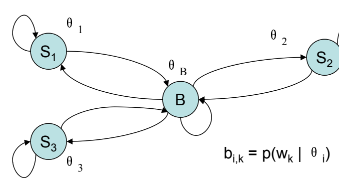

Figure 4: A 3-theme HMM

The background state, which corresponds to the background theme model, aims to account for non-discriminative words, while the content words and subtopics are modeled by the states corresponding to the trans-collection themes. Since the extracted themes are discriminative, we may reasonably assume that each theme can only shift to another theme through the background model. The unknown parameter set in the HMM is $\Lambda = \{\pi_i, a_{i,i}, a_{i,B}, a_{B,i}\}_{i=1}^n$. Λ can be estimated using an EM algorithm called Baum-Welch algorithm [17].

After the initial state probabilities and transition probabilities are estimated, the Viterbi algorithm [17] can be used to decode the text stream to obtain the most likely state sequence, i.e., the most likely sequence of theme shifts, as shown in Figure 5.

Once the whole stream $C = \{d_1, ..., d_T\}$ is decoded with the labels of themes, we can use a fixed-size sliding window of time to measure the strength of each theme at a time point[1]. Let $d_i = d_{i1}...d_{i|d_i|}$ be the sequence of words in d_i. The absolute and relative strengths of theme i at time t is computed as:

$$AStrength(i,t) = \frac{1}{W}\sum_{t'\in[t-\frac{W}{2},t+\frac{W}{2}]}\sum_{j=1}^{|d_{t'}|}\delta(d_{t'j}, i)$$

[1]The use of a sliding window also avoids the "report delay" problem in the news domain.

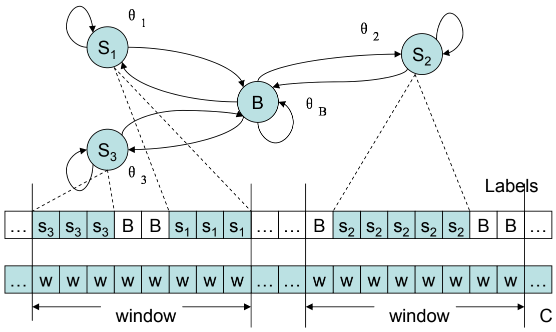

Figure 5: Decoding the collection

where $\delta(d_{t'j}, i) = 1$ if word $d_{t'j}$ is labeled as theme i; otherwise $\delta(d_{t'j}, i) = 0$. W is the size of the sliding window in terms of time points.

$$NStrength(i,t) = \frac{AStrength(i,t)}{\sum_{j=1}^{k} AStrength(j,t)} = \frac{\sum_{t' \in [t-\frac{W}{2}, t+\frac{W}{2}]} \sum_{j=1}^{|d_{t'}|} \delta(d_{t'j}, i)}{\sum_{t' \in [t-\frac{W}{2}, t+\frac{W}{2}]} |d_{t'}|}$$

The life cycle of each theme can then be modeled as the variation of the theme strengths over time.

The analysis of theme life cycles thus involves the following four steps: (1) Construct an HMM to model how themes shift between each other in the collection. (2) Estimate the unknown parameters of the HMM using the whole stream collection as observed example sequence. (3) Decode the collection and label each word with the hidden theme model from which it is generated. (4) For each trans-collection theme, analyze when it starts, when it terminates, and how it varies over time.

5. EXPERIMENTS AND RESULTS

5.1 Data Preparation

Two data sets are constructed to evaluate the proposed ETP discovery methods. The first, tsunami news data, consists of news articles about the event of Asia Tsunami dated Dec. 19 2004 to Feb. 8 2005. We downloaded 7468 news articles from 10 selected sources, with the keyword query "tsunami". As shown in Table 1, three of the sources are in Asia, two of them are in Europe and the rest are in the U.S.

News Source	Nation	News Source	Nation
BBC	UK	Times of India	India
CNN	US	VOA	US
Economics Times	India	Washington Post	US
New York Times	US	Washington Times	US
Reuters	UK	Xinhua News	China

Table 1: News sources of Asia Tsunami data set

The second data set consists of the abstracts in KDD conference proceedings from 1999 to 2004. All the abstracts were extracted from the full-text pdf files downloaded from the ACM digital library[2]. 2 articles were excluded because they were not recognizable by the pdf2text software in Linux, giving us a total of 496 abstracts. The basic statistics of the two data sets are shown in Table 2. We intentionally did not perform stemming or stop word pruning in order to test the robustness of our algorithms.

Data Set	# of docs	AvgLength	Time range
Asia Tsunami	7468	505.24	12/19/04 - 02/08/05
KDD Abs.	496	169.50	1999-2004

Table 2: Basic information of data sets

On each data set, two experiments are designed: (1) Partition the collection into time intervals, discover the theme evolution graph and identify theme evolution threads. (2) Discover trans-collection themes and analyze their life cycles. The results are discussed below.

5.2 Experiments on Asia Tsunami

Since news reports on the same topic may appear earlier in one source but later in another (i.e., "report delay"), partitioning news articles into *overlapping*, as opposed to non-overlapping subcollections seems to be more reasonable. We thus partition the our news data into 5 time intervals, each of which spans about two weeks and is half overlapping with the previous one. We use the mixture model discussed in Section 3 to extract the most salient themes in each time interval. We set the background parameter $\lambda_B = 0.95$ and number of themes in each time interval to be 6. The variation of λ_B is discussed later. Table 3 shows the top 10 words with the highest probabilities in each theme span. We see that most of these themes suggest meaningful subtopics in the context of the Asia tsunami event.

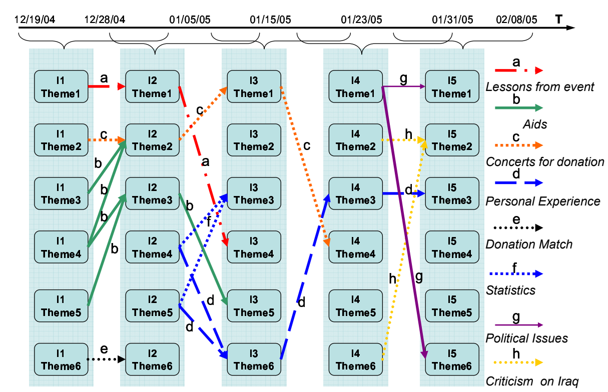

Figure 6: Theme evolution graph for Asia Tsunami

With these theme spans, we use KL-divergence to further identify evolutionary transitions. Figure 6 shows a theme evolution graph discovered from Asia Tsunami data when the threshold for evolution distance is set to $\xi = 12$. From Figure 6, we can see several interesting evolution threads which are annotated with symbols.

The thread labeled with a may be about warning systems for tsunami. It is interesting to see that the nation covered by the thread seems to have evolved from the U.S. in period l_1, to China in l_2, and then to Japan in l_3. In thread b, themes 3, 4, and 5 in period l_1 indicate the aids and financial support from UN, from local area, and special aids for children, respectively. They all show an evolutionary transition to theme 2 (donation from UK) and theme 3 (aid from

[2]http://www.acm.org/dl

	Theme 1	Theme 2	Theme 3	Theme 4	Theme 5	Theme 6
l_1:	system 0.0104	Year 0.0074	debt 0.0148	Aceh 0.0320	Annan 0.0081	match 0.0094
Dec/	Bush 0.0080	silence 0.0056	Club 0.0098	Indonesia 0.0118	U.N. 0.0064	XI 0.0065
19/	warning 0.0070	British 0.0053	Paris 0.0097	military 0.0118	summit 0.0062	players 0.0059
04	dollars 0.0067	New 0.0051	Bank 0.0063	Banda 0.0096	children 0.0060	Cricket 0.0058
	million 0.0064	celebrations 0.0050	moratorium 0.0061	Indonesian 0.0089	Powell 0.0044	game 0.0050
-	small 0.0058	UK 0.0047	freeze 0.0058	province 0.0088	NBC 0.0037	Zealand 0.0044
	US 0.0055	music 0.0038	repayments 0.0052	workers 0.0087	million 0.0036	Australia 0.0042
Jan/	conference 0.0052	London 0.0038	billion 0.0044	foreign 0.0081	disease 0.0035	Sudan 0.0039
04/	meeting 0.0035	Sydney 0.0037	U.N. 0.0044	islands 0.0077	WHO 0.0033	captain 0.0038
05	Egeland 0.0033	Blair 0.0035	nations 0.0042	aid 0.0071	UNICEF 0.0031	Ponting 0.0036
l_2:	countries 0.0240	Mr 0.0104	Aceh 0.0226	missing 0.0143	her 0.0147	match 0.0075
Dec/	debt 0.0146	Blair 0.0068	aid 0.0204	Thailand 0.0115	islands 0.0102	Cricket 0.0065
28/	system 0.0085	British 0.0062	Powell 0.0171	bodies 0.0107	Nicobar 0.0098	players 0.0052
04	nations 0.0084	Rs 0.0057	relief 0.0161	dead 0.0071	I 0.0069	XI 0.0052
	China 0.0073	Britons 0.0047	Indonesia 0.0160	Sweden 0.0068	she 0.0067	you 0.0046
-	warning 0.0064	UK 0.0046	Annan 0.0134	Thai 0.0065	Andaman 0.0064	Zealand 0.0042
	Paris 0.0064	donations 0.0045	U.S. 0.0131	Swedish 0.0064	beach 0.0064	game 0.0033
Jan/	Club 0.0058	crore 0.0037	United 0.0122	police 0.0060	sea 0.0064	points 0.0033
14/	Bank 0.0056	Tamil 0.0036	military 0.0113	DNA 0.0056	my 0.0060	captain 0.0032
05	Chinese 0.0054	public 0.0033	U.N. 0.0110	tourists 0.0052	island 0.0051	cricket 0.0032
l_3:	Chinese 0.0085	Tamil 0.0121	toll 0.0103	warning 0.0121	United 0.0228	Thailand 0.0103
Jan/	British 0.0076	Sri 0.0121	bodies 0.0083	system 0.0119	Powell 0.0168	missing 0.0092
05/	UK 0.0075	Lanka 0.0070	death 0.0067	islands 0.0086	Bush 0.0165	Phuket 0.0087
05	China 0.0070	Nadu 0.0061	dead 0.0067	sea 0.0061	U.S. 0.0146	Khao 0.0070
	Hong 0.0068	Tigers 0.0059	debt 0.0063	Nicobar 0.0048	States 0.0137	her 0.0070
-	Kong 0.0064	government 0.0050	food 0.0057	Pacific 0.0047	Mr. 0.0117	beach 0.0068
	donations 0.0060	Lankan 0.0040	Paris 0.0057	water 0.0042	Nations 0.0101	Lak 0.0067
Jan/	Red 0.0056	Nicobar 0.0040	Indonesia 0.0056	Japan 0.0040	$ 0.0088	Swedish 0.0066
22/	concert 0.0052	Singh 0.0037	Club 0.0053	Kobe 0.0037	relief 0.0079	Sweden 0.0064
05	Cross 0.0050	rebels 0.0031	corpses 0.0051	quake 0.0033	million 0.0076	hotel 0.0059
l_4:	Aceh 0.0250	funding 0.0046	Phi 0.0052	concert 0.0107	LTTE 0.0055	Iraq 0.0087
Jan/	talks 0.0175	Iraq 0.0044	her 0.0048	Kobe 0.0050	Tamil 0.0052	Bush 0.0086
15/	GAM 0.0150	Eid 0.0039	ASEAN 0.0036	singer 0.0045	talks 0.0037	billion 0.0084
05	rebels 0.0133	regional 0.0035	resort 0.0024	stars 0.0041	local 0.0036	pilgrims 0.0073
	peace 0.0100	festival 0.0034	Palu 0.0023	Stadium 0.0040	UK 0.0034	budget 0.0067
-	Indonesian 0.0085	congressional 0.0033	Palu 0.0023	Wales 0.0040	Tigers 0.0033	deficit 0.0060
	province 0.0074	mosque 0.0033	cancer 0.0022	Japan 0.0036	Hafun 0.0030	House 0.0059
Jan/	Free 0.0055	Rice 0.0032	Phuket 0.0021	rock 0.0035	Norwegian 0.0030	boat 0.0053
30/	Movement 0.0052	month 0.0030	Hui 0.0021	Millennium 0.0034	Prabhakaran 0.0029	Trump 0.0042
05	rebel 0.0048	military 0.0029	Fleming 0.0020	Live 0.0030	Kalpakkam 0.0028	spending 0.0042
l_5:	Jones 0.0051	billion 0.0197	boat 0.0081	Clinton 0.0115	debt 0.0195	talks 0.0263
Jan/	Palu 0.0046	$ 0.0153	tourism 0.0067	var 0.0052	meeting 0.0136	Aceh 0.0213
23/	station 0.0045	Iraq 0.0140	Samui 0.0059	Nepal 0.0049	finance 0.0122	peace 0.0147
05	Pierson 0.0042	House 0.0121	ASEAN 0.0055	summit 0.0044	Brown 0.0087	Indonesian 0.0113
	song 0.0034	budget 0.0101	JAL 0.0054	SAARC 0.0042	exchange 0.0074	rebels 0.0112
-	North 0.0033	request 0.0094	tourists 0.0046	Dhaka 0.0036	ministers 0.0067	Helsinki 0.0094
	Korea 0.0033	funding 0.0086	accident 0.0041	construction 0.0030	agreed 0.0065	conflict 0.0077
Feb/	Miss 0.0031	White 0.0083	month 0.0041	Bangladesh 0.0026	gold 0.0054	province 0.0070
8/	97 0.0030	Afghanistan 0.0071	joke 0.0038	envoy 0.0025	IMF 0.0054	sides
05	show 0.0030	baby 0.0066	Marsh 0.0035	techniques 0.0021	economic 0.0047	autonomy

Table 3: Theme spans extracted from Asia Tsunami data

US) in l_2. The latter theme further evolves into theme 5 in l_3, which is mainly about money support from US. Thread c begins with music-related events and aids from UK. It shifts to talk about concerts in Hong Kong and then Japan with the purpose of raising funds for donation. Thread d is about the personal experiences of survivors. It starts with theme 5 in l_2, goes through theme 6 in l_3, theme 3 in l_4, and finally evolves into theme 3 in l_5. There are also several short but noticeable theme evolution threads. For example, thread e is about cricket matches for donation, while thread f is about deaths and losses in the disaster.

In the latest two time intervals, most themes are no longer about the tsunami event, indicating that the event was probably receiving diminishing attention in these two periods, which can be seen more clearly later from the analysis of the life cycles of themes. There are two politics-related short theme threads (i.e., g and h). In thread g, theme 1 in l_4 is about political issues ("rebels" and "peace"). It splits into two themes in l_5, about North Korea and the Aceh peace talk, respectively. Theme 2 and theme 5 in l_4 represent criticisms on the Iraq affair (one for military issues and one for the high expenditure/cost). In l_5, they merged into a single theme, which mentions the budget on Iraq and Afghanistan issues. Interestingly, by linking back to the articles, it turns out to be arguing for shrinking the budget on the war issues and offering more aid for the disaster.

Note that multiple threads may share one or more common themes, resulting in thread ambiguity. For example, themes 2, 3 and 4 of l_1 all have a high similarity to theme 2 of l_2. In the analysis above, we only included theme 2 of l_1 in thread c, because themes 3 and 4 do not appear to be similar to theme 2 of l_2 in the same way as theme 1 of l_3 is. A very interesting future research direction would be to study how we can automatically perform thread disambiguation.

Our second experiment aims to model the life cycles of trans-collection themes. In this experiment, we use two individual sources (CNN and Xinhua News) instead of the whole mixed collection to avoid "report delay". The five trans-collection themes extracted from CNN and Xinhua News are shown in Table 4.

The five themes from CNN roughly correspond to (1) research and lessons about the tsunami; (2) personal experience of survivors; (3) Special aid program for children; (4) general reports and statistics; (5) aids and donations from the world, especially from the U.S. The five themes from Xinhua roughly correspond to (1) statistics of death and missing; (2) reports and stories at the scene; (3) donations from China; (4) aids and donations from the world; (5) research and lessons about the tsunami. Some themes (e.g., CNN-theme1 and XINHUA-theme5) are common to both sources, while some others (e.g., CNN-theme5 and XINHUA-theme3) clearly reflect the different regions of the two sources.

In Figure 7 we plot the absolute strengths of the trans-collection themes over time for CNN ($W = 10$). We see that the absolute strengths of all five themes are increasing in the first 10 days after Dec. 24, 2004. Reports on aids for children and aids from the world begin to decay after that. General reports and statistics starts to decay around Jan 10 for the rest of the time. Around Jan. 7th, the theme

Source	Theme 1	Theme 2	Theme 3	Theme 4	Theme 5
	system 0.0079	I 0.0322	children 0.0119	Aceh 0.0088	Bush 0.0201
	warning 0.0075	wave 0.0061	debt 0.0072	Indonesia 0.0063	$ 0.0173
C	Ocean 0.0073	beach 0.0056	hospital 0.0072	said 0.0054	million 0.0135
N	Indian 0.0064	water 0.0051	baby 0.0064	military 0.0044	relief 0.0134
N	Pacific 0.0063	when 0.0050	Club 0.0063	U.N. 0.0038	United 0.0105
	earthquake 0.0061	saw 0.0046	Paris 0.0061	number 0.0032	aid 0.0099
	quake 0.0057	sea 0.0046	child 0.0054	survivors 0.0032	Powell 0.0098
	tsunami 0.0054	Thailand 0.0042	her 0.0053	reported 0.0031	U.S. 0.0075
	ocean 0.0039	family 0.0039	police 0.0048	helicopters 0.0028	States 0.0075
	scientists 0.0031	ran 0.0033	moratorium 0.0046	killed 0.0027	U.N. 0.0056
	Thailand 0.0104	Aceh 0.0219	Chinese 0.0391	dollars 0.0226	system 0.0314
X	Thai 0.0096	province 0.0111	China 0.0391	million 0.0204	warning 0.0272
I	missing 0.0079	Indonesian 0.0075	yuan 0.0180	US 0.0178	early 0.0172
N	victims 0.0054	tidal 0.0055	countries 0.0098	aid 0.0118	meeting 0.0159
H	Philippine 0.0040	waves 0.0047	Beijing 0.0089	United 0.0108	Ocean 0.0121
U	confirmed 0.0040	killed 0.0045	travel 0.0061	countries 0.0106	small 0.0096
A	residents 0.0037	quake 0.0043	$ 0.0058	UN 0.0102	international 0.0092
	tourists 0.0033	island 0.0043	donated 0.0057	Annan 0.0082	conference 0.0086
	percent 0.0032	dead 0.0041	Cross 0.0053	debt 0.0071	natural 0.0082
	number 0.0032	death 0.0041	donation 0.0052	reconstruction 0.0062	disasters 0.0070

Table 4: Trans-collection themes extracted from CNN and Xinhua News

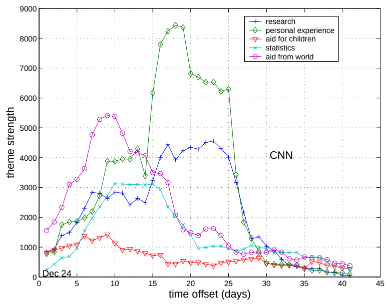

Figure 7: Absolute strength life cycles in CNN data

on the research and lessons about tsunami starts to increase again. The same pattern is discovered in reports on personal experiences, which is probably because survivors had come back to their home country around that time. Both themes drop sharply around Jan. 17. After Jan. 22, all 5 themes retain a low strength level, indicating the event was receiving diminishing attention. The normalized strengths of themes in the CNN data show similar patterns.

In Figure 8, we show the absolute and normalized strengths of the five trans-collection themes over time in Xinghua News ($W = 10$). We see that, in the first week beginning Dec. 25th 2004, all 5 themes are increasing rapidly, but they all begin to decay around Jan. 10th except for stories and reports at the scene, which increases again after a roughly 10-day period of mild decreasing. The theme about death statistics begins to decay all the time after Jan. 16. Both aids from China and the research and lessons about tsunami present a second rise in late January, although not as significant as the first one. In the normalized strength plot, it is easy to see that before Jan 3rd, the dominating theme is theme 5. In the next 10 days, aid from the world is most significant. In the following 20 days, "on-scene stories" is the dominating theme, although its absolute strength is decreasing for most of the time. In the last time period when the overall coverage of the topic had significantly decreased, Aids from China is relatively stronger than other themes. Comparing CNN and Xinhua, we see the life cycles of the correlated themes in the two data sets exhibit comparable patterns but with some differences.

5.3 Experiments on KDD Abstracts

The publication year naturally suggests a non-overlapping partition of the KDD abstract data. We thus treat all the abstracts published in one year as one time interval. The theme spans extracted from each year using the mixture

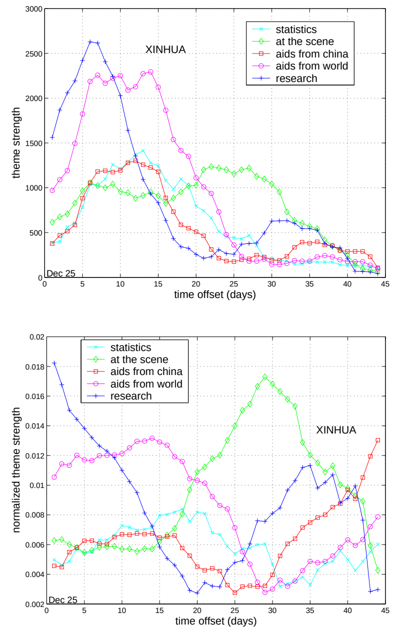

Figure 8: Absolute (top) and normalized (bottom) strength life cycles in Xinhua News

model with $\lambda_B = 0.9$ are shown in Table 5. The number of themes slightly differs from year to year because we apply a threshold to select only the most salient themes as described in Section 3. Similar to what we have seen on the news data, the themes here are also mostly meaningful in the context of KDD publications. The three themes in the year of 1999 are about association rule mining, clustering, and classification respectively, which are all traditional data mining topics, compared with the new topics, such as spatial data mining (theme 1) and gene and microarray mining (theme 2), extracted in the year of 2004.

A theme evolution graph extracted using an evolution distance threshold of $\xi = 12.5$ is shown in Figure 9, where we see several interesting theme threads.

Thread a starts with theme 3 in 1999 (about classification). It first evolves into theme 1 in 2001 (typical classification techniques such as SVM), and then evolves into web classification in 2002. The next theme span on this thread is about clustering and random variables, which has some influence on theme 1 in 2004. Another evolution thread (b) starts with association rule mining in 1999, and transits into frequent item set in 2001. Both themes show strong evolu-

	1999	2000	2001	2002	2003	2004
THEME 1	association 0.0156	ABSTRACT 0.0141	rules 0.0148	web 0.0089	Clustering 0.0077	topic 0.0104
	rules 0.0149	revision 0.0121	datasets 0.0104	hierarchical 0.0081	indices 0.0077	Algorithms 0.0103
	associations 0.0128	clustering 0.0120	artificial 0.0086	classification 0.0068	Recognition 0.0070	correlation 0.0086
	rule 0.0090	test 0.0106	support 0.0070	features 0.0057	mixture 0.00654	image 0.0079
	attribute 0.0062	die 0.0094	rule 0.0068	analysis 0.0055	Pattern 0.0065	mixture 0.0076
	dimension 0.0060	Terms 0.0093	SVM 0.0067	Markov 0.0053	random 0.0065	LDA 0.0060
	median 0.0057	classi 0.0090	criteria 0.0063	systems 0.0050	cluster 0.0060	metrics 0.0055
	attributes 0.0056	protein 0.0077	classification 0.0062	topics 0.0046	components 0.0055	spatial 0.0048
	polish 0.0050	control 0.0076	linear 0.0061	classification. 0.0046	clustering 0.0053	semantic 0.0048
	transaction 0.0047	functional 0.0060	useful 0.0061	intrusion 0.0045	variables 0.0052	incremental 0.0048
THEME 2	algorithms 0.0155	rules 0.0192	disconnected 0.0102	gene 0.0164	Information 0.0122	genes 0.0184
	Abstract 0.0084	function 0.0116	web 0.0097	time 0.0153	cube 0.0117	problem 0.0100
	EPs 0.0077	sequence 0.0091	models 0.0092	series 0.0139	Web 0.0096	graph 0.0099
	objects 0.0059	set 0.0071	components 0.0083	change 0.0106	social 0.0075	structure 0.0088
	distance 0.0054	minimality 0.0067	graph 0.0082	statistical 0.0072	weighted 0.0066	Algorithms 0.0076
	clustering 0.0052	discovered 0.0064	boosting 0.0070	may 0.0069	Retrieval 0.0065	collection 0.0071
	sampling 0.0050	protein 0.0063	dimensionality 0.0068	detection 0.0065	distance 0.0059	subset 0.0068
	problem 0.0048	minimal 0.0061	random 0.0065	source 0.0063	user 0.0059	microarray 0.0060
	Bayesian 0.0046	functional 0.0060	reduction 0.0056	kernel 0.0062	Search 0.0054	semantic 0.0059
	profiles 0.0044	tuberculosis 0.0056	outlier 0.0052	base 0.0061	networks 0.0047	samples 0.0057
THEME 3	Abstract 0.0104	students 0.0110	item 0.0117	classification 0.0150	Database 0.0102	part 0.0120
	itemsets 0.0085	level 0.0088	database 0.0096	algorithm 0.0130	data 0.0091	reviews 0.0083
	rules 0.0084	aggregate 0.0083	sets 0.0091	text 0.0128	expression 0.0082	BOM 0.0076
	local 0.0068	statistical 0.0081	frequent 0.0065	unlabeled 0.0113	gene 0.0074	opinion 0.0070
	tree 0.0067	statistics 0.0077	compounds 0.0059	document 0.0087	results 0.0069	maximal 0.0070
	decision 0.0063	user 0.0075	unexpected 0.0055	documents 0.0076	cabin-level 0.0069	sentences 0.0060
	classifier 0.0062	cation 0.0072	knowledge 0.0054	labeled 0.0075	mining 0.0068	product 0.0058
	class 0.0061	distributed 0.0071	changes 0.0051	customer 0.0074	voting 0.0068	receiver 0.0053
	incremental 0.0056	decision 0.0070	MOLFEA 0.0050	approach 0.0073	study 0.0066	positive 0.0049
	Bayes 0.0053	models 0.0067	human 0.0049	learning 0.0067	tree 0.0055	internal 0.0044
			(2002 THEME5)			
THEME 4		manufacturing 0.0122	(rule 0.0250)	clustering 0.0070	copies 0.0063	inference 0.0071
		problems 0.0099	(optimal 0.0094)	retrieval 0.0065	stream 0.0059	failing 0.0068
		demographic 0.0094	(clustering 0.0078)	SyMP 0.0058	MLC 0.0056	Learning 0.0061
		customers 0.0081	(ROCCH 0.0075)	estimators 0.0057	generation 0.0052	TiVo 0.0059
		training 0.0079	(smoothing 0.0073)	very 0.0054	Data 0.0051	itemsets 0.0059
		similarity 0.0074	(association 0.0068)	complexity 0.0053	2003 0.0047	shopping 0.0058
		yield 0.0070	(sets 0.0066)	different 0.0049	image 0.0047	Pattern 0.0057
		SVMs 0.0065	(causality 0.0065)	behavior 0.0049	frequent 0.0044	image 0.0056
		cost 0.0062	(corpus 0.0064)	each 0.0049	part 0.0041	dense 0.0055
		measures 0.0058	(term 0.0064)	small 0.0047	important 0.0041	constraints 0.0052

Table 5: Theme spans extracted from KDD Abstract Data

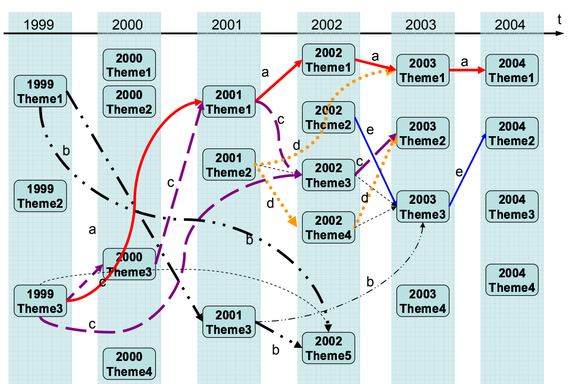

Figure 9: Theme evolution graph from KDD Data

tionary relations to theme 5 in 2002, and the frequent item set theme shows some weak evolutionary influence on theme 3 in 2003 (gene expression mining).

Another interesting group of edges are tagged with symbol c. Starting with classification in 1999, it transits into theme 3 in 2000 (statistical analysis and decision making). The path further connects theme 1 in 2001, and then makes an interesting transition into text classification in 2002. Theme 3 in 1999 itself also has a strong evolutionary relation to text classification. This path ends at theme 2 in 2003, which covers the Web and social networks. The discussion of web mining has not appeared as an explicit theme until theme 2 in 2001, which evolves into web and social networks in 2003 through theme 1 in 2002.

Before year 2000, there was no explicit theme about data mining in biological data. There are two themes (theme 1, theme 2) which mention protein functionalities, but they fail to reappear in 2001. A strong explicit theme evolution thread of mining biological data (path e) starts with theme 2 in year 2002 and evolves into theme 3 in 2003 and theme 4 in 2004 , respectively.

As in the news data, we also analyzed the life cycles of trans-collection themes in KDD Abstracts. Seven dominating trans-collection themes are shown in Table 6 and the interesting patterns of life cycles are presented in Figure 10 ($W = 1$). Some new topics, such as spatial-temporal data mining, have not shown up as trans-collection themes, because when we consider the whole stream, they are not among the dominating topics. From Figure 10, we see that the normalized strength of theme 1 decreases monotonically from 1999. This theme is about a traditional application area of data mining – marketing and customer analysis. Another theme showing a decaying pattern is association rule mining, which keeps decreasing after its peak in 2000. In the year 1999, there is very little work on mining web information. This topic keeps growing in the following three years, and drops a bit after its acme in the year 2002. Mining from genes and biology data, as highlighted, keeps increasing over the 6 years from a very low level to one of the strongest themes. Theme 4, which covers time series and other applications of clustering, shows an irregular pattern before 2002 but remains stable after that.

There are also themes, such as classification (theme 5) and clustering (theme 7, mostly theoretical aspect, especially dimension reduction), which are somehow stable. Indeed, the classification theme appears to be among the strongest themes over the whole time line. Considering that several themes (theme 3, theme 4, and theme 7) all cover clustering, we may also infer that clustering is another major dominating theme in KDD publications.

5.4 Parameter Tuning

In our methods for ETP discovery, there are a few parameters that are meant to provide the necessary flexibility for a

Theme 1	Theme 2	Theme 3	Theme 4	Theme 5	Theme 6	Theme 7
marketing 0.0087	gene 0.0173	clustering 0.0082	set 0.0076	tree 0.0107	rules 0.0142	distance 0.0150
customer 0.0086	genes 0.0104	Web 0.0070	series 0.0076	decision 0.0094	association 0.0064	objects 0.0094
models 0.0079	expression 0.0096	selection 0.0064	manufacturing 0.0066	classification 0.0086	support 0.0063	reduction 0.0071
customers 0.0076	probability 0.0081	user 0.0056	time 0.0065	itemsets 0.0061	rule 0.0060	clustering 0.0061
business 0.0048	time 0.0063	text 0.0050	clustering 0.0065	Bayes 0.0057	framework 0.0050	similarity 0.0052
no-show 0.0042	coherent 0.0058	distance 0.0048	test 0.0056	estimates 0.0052	outliers 0.0048	Euclidean 0.0050
Web 0.0041	microarray 0.0038	pages 0.0040	patterns 0.0050	probability 0.0052	useful 0.0040	dimensionality 0.0043

Table 6: Trans-collection themes extracted from KDD Abstract Data

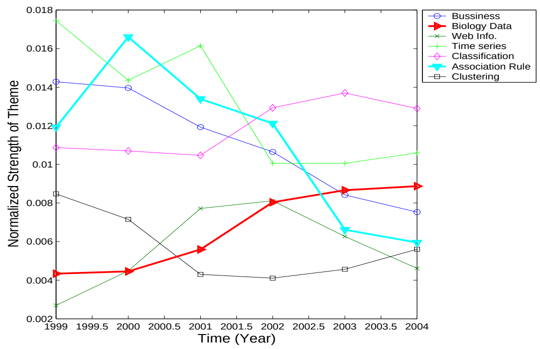

Figure 10: Normalized strength life cycles in KDD

user to control the pattern analysis results. We now discuss them in some detail.

In the mixture model for theme extraction, λ_B controls the strength of our background model, which is used to "absorb" non-informative words from the themes. In general, λ_B should be set to reflect a user's knowledge about the noise (i.e., the non-informative common words) in the text stream; the more noise we believe our data set has, the larger λ_B should be. Our experiments have shown that, in ordinary English text, the value of λ_B can be set to a value between 0.9 and 0.95. Within this range, the setting of λ_B does not affect the extracted themes significantly, but it does affect the top words with the highest probabilities; a smaller λ_B tends to cause non-informative common words to show up in the top word list. Parameter k represents the expected number of subtopics in a subcollection. In our experiments, we determine the number of themes by using a relatively large k and drop a theme j, if the value of $\frac{1}{|C_i|}\sum_{d\in C_i}\pi_{d,j}$ is significantly low.

Another parameter is the evolution distance threshold ξ. This parameter has a "zooming" effect for the discovered theme evolution graph. A tighter (smaller) ξ would only allow us to see the strongest evolutionary transitions, whereas a looser (larger) ξ would allow us to examine some weaker evolutionary transitions as well.

Yet another parameter is the size of sliding window W, which controls the amount of supporting documents when computing the strength of theme θ at time t and affects the smoothness of the life cycle curve. A small W introduces less smoothing and would allow us to see the temporal patterns in high resolution, but may also make it difficult to see the overall trend. A larger W would give a smoother curve, but may hide some interesting local variation patterns. Regarding the "report delay" problem in the news domain, a reasonable value for W appears to be 7-15 days (3-7 days at each side of time t).

6. RELATED WORK

While TTM has not been well studied, there are several lines of research related to our work. For example, in Kleinberg's work on discovering bursty and hierarchical structures in streams [10], text streams are converted to temporal frequency data and an infinite-state automaton is used to model the stream. Detection of novel topics and trends in text streams has been studied by several researchers [3, 18, 19, 11, 13, 14], but their focus is to identify emerging trends rather than summarize the complete evolutionary theme patterns in a given text stream as we do.

An interesting related work to our analysis of theme life cycles is [16], where Perkio and others used a Multinomial PCA model to extract themes from a text collection and they used a hidden theme-document weight, which is similar to $\pi_{d,j}$ in Section 3, to compute the strength of a theme. The major difference between our work and theirs is that we model the theme transitions in a context-sensitive way with an HMM, which presumably captures the natural proximity of similar topics better.

Text clustering is another well studied problem relevant to our work. Specifically, the aspect models studied in [9, 20, 2] are related to the mixture theme model we use to extract themes. However, these works do not consider temporal structures in text. Nallapati and others studied how to discover sub-clusters in a news event and structure them by their dependency, which could also generate a graph structure [15]. A major difference between our work and theirs is that they perform document level clustering, while we perform theme level word clustering. Another difference is that they do not consider the variations of subtopics in different time periods while we analyze life cycles of themes.

Since a theme evolution graph and theme life cycle can serve as a good summary of a collection, our work is also partially related to document summarization (e.g., [12, 1]). Allan and others presented a news summarization method based on ranking and selecting sentences obeying temporal order [1]. However, summarization intends to retain the explicit information in text in order to maintain fidelity, while we aim at extracting non-obvious implicit themes and their evolutionary patterns.

7. CONCLUSIONS

Text streams often contain latent temporal theme structures which reflect how different themes influence each other and evolve over time. Discovering such evolutionary theme patterns can not only reveal the hidden topic structures, but also facilitate navigation and digestion of information based on meaningful thematic threads. In this paper, we propose general probabilistic approaches to discover evolutionary theme patterns from text streams in a completely unsupervised way. To discover the evolutionary theme graph, our method would first generate word clusters (i.e., themes) for each time period and then use the Kullback-Leibler divergence measure to discover coherent themes over time. Such an evolution graph can reveal how themes change over time and how one theme in one time period has influenced other themes in later periods. We also propose a method based on hidden Markov models for analyzing the life cycle of each theme. This method would first discover the globally interesting themes and then compute the strength of a theme in each time period. This allows us to not only see the trends of strength variations of themes, but also compare the relative strengths of different themes over time.

We evaluated our methods using two different data sets. One is a stream of 50 days' news articles about the tsunami disaster that happened recently in Asia, and the other is the abstracts of the KDD conference proceedings from 1999 to 2004. In both cases, the proposed methods can generate meaningful temporal theme structures and allow us to summarize and analyze the text data from temporal perspective. Our methods are generally applicable to any text stream data and thus have many potential applications in temporal text mining.

There are several interesting directions to further extend this work. First, we have only considered a flat structure of themes; it would be interesting to explore hierarchical theme clustering, which can give us a picture of theme evolutions at different resolutions. Second, we can develop a temporal theme mining system based on the proposed methods to help a user navigate the stream information space based on evolutionary structures of themes. Such a system can be very useful for managing all kinds of text stream data. Finally, temporal text mining (TTM) represents a promising new direction in text mining that has not yet been well-explored. In addition to evolutionary theme patterns, there are many other interesting patterns such as associations of themes across multiple streams that are interesting to study.

8. ACKNOWLEDGMENTS

We thank Tao Tao for his constructive technical comments and Hang Su for helping collect the tsunami data. We are grateful to the three anonymous reviewers for their extremely useful comments. This material is based in part upon work supported by the National Science Foundation under award numbers CAREER-IIS-0347933 and ITR-IIS-0428472.

9. REFERENCES

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of news topics. In *Proceedings of ACM SIGIR 2001*, pages 10–18, 2001.

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

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

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

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

[6] 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.

[7] R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In *KDD*, pages 112–117, 1995.

[8] M. A. Hearst. Untangling text data mining. In *Proceedings of the 37th conference on Association for Computational Linguistics (ACL 1999)*, pages 3–10, 1999.

[9] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[10] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101, 2002.

[11] 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.

[12] R. Kumar, U. Mahadevan, and D. Sivakumar. A graph-theoretic approach to extract storylines from search results. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 216–225, 2004.

[13] J. Ma and S. Perkins. Online novelty detection on temporal sequences. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–618, 2003.

[14] S. Morinaga and K. Yamanishi. Tracking dynamics of topic trends using a finite mixture model. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 811–816, 2004.

[15] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *Proceedings of the Thirteenth ACM conference on Information and knowledge management*, pages 446–453, 2004.

[16] J. Perkio, W. Buntine, and S. Perttu. Exploring independent trends in a topic-based search engine. In *Proceedings of the Web Intelligence, IEEE/WIC/ACM International Conference on (WI'04)*, pages 664–668, 2004.

[17] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–285, Feb. 1989.

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

[19] S. Roy, D. Gevry, and W. M. Pottenger. Methodologies for trend detection in textual data mining. In *the Textmine '02 Workshop, Second SIAM International Conference on Data Mining*, 2002.

[20] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 743–748, 2004.