

Discovering Evolutionary Theme Patterns from Text*

CS598CXZ Project Report, Professor ChengXiang Zhai

Qiaozhu Mei
qmei2@uiuc.edu

May 5, 2005

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. The evolutionary patterns of a theme are divided into content evolution and strength evolution. We define the problem of discovering evolutionary theme patterns on both aspects and present general probabilistic methods for (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 three different data collections (i.e., one news articles collection and two literature collections) shows that the proposed methods can discover interesting evolutionary theme patterns effectively.

1 Introduction

In many application domains, we encounter a stream of text data, 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 where the articles are ordered according to their publication dates. 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 of the event, the progression of the event, and its impact, 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 tsunami disaster that happened in Asia in the end of 2004. A query to <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, which includes not only the subtopics about this event, some threads corresponding to the evolution of these themes, would be extremely useful. 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

*An alternative paper about this project is submitted to KDD 2005.

subsequent themes it influences. A timeline based theme structure as shown in would be a very informative summary of the event, which provides threads supporting navigation of themes.

Now consider another scenario in the scientific literature domain. In a given research area, there are often hundreds of papers published annually. It would be very useful for a researcher, especially a beginning researcher, to understand the evolution of research topics in the literature. 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, 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.

These different needs have revealed two major categories of evolutionary theme patterns: (1) evolution of theme contents; and (2) evolution of theme strength. 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 two categories of evolutionary theme patterns. Since most information bears some kinds of time stamps, we may expect ETP discovery to have many applications in multiple domains, e.g. email analysis, mining user logs, mining customer reviews, in addition to news summarization and literature mining.

Despite its importance, however, to the best of our knowledge, TTM has not been well addressed in the existing work. Most existing text mining work [3, 4] does not consider the temporal structures of text. Kleinberg’s work on discovering bursty and hierarchical structures in streams [5] represents a major previous work on TTM, in which text streams are converted to temporal frequency data and an infinite-state automaton is used to model the stream. However, this method is inadequate for generating the evolutionary theme patterns as shown in the two examples above.

The rest of the paper is organized as follows. In section 2, we will introduce the methodology of this project. Experiments on three different data sets are presented in section 3. Further discussions will follow in section 4.

2 Methodology

Given a stream collection of text C , a major task of a general **Evolutionary Theme Pattern (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. 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 major tasks of ETP discovery include: (a) extracting theme sects at different time intervals; (b) discovering evolutionary transitions between theme sects and discovering theme evolution threads over time; (c) modeling the life cycles of trans-collection themes.

2.1 Evolution Graph Discovery

Given a stream of text $C = \{d_1, d_2, \dots, d_T\}$, our goal is to extract a theme evolution graph from C automatically. As in many other text mining tasks, we generally rely on unsupervised learning techniques such as clustering to discover the theme sects and their evolution relations. Specifically, we follow the following procedure:

1. Partition the documents into n possibly overlapping subcollections with a fixed or variable time intervals so that $C = C_1 \cup \dots \cup C_n$ and $C_i = \{d_{t_i}, \dots, d_{t_i+l_i-1}\}$ is a subcollection of l documents in the time

span $[t_i, t_i + l_i - 1]$. In general, $t_i < t_{i+1}$, but it may be that $t_i + l_i - 1 > t_{i+1}$, so that 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 $\Theta_i = \{\theta_{i,1}, \dots, \theta_{i,k_i}\}$ 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 .

2.2 Theme Extraction

We extract themes from each subcollection C_i using the simple probabilistic mixture model presented in [7]. The basic idea of this method is to treat the words as observations from a mixture model where the component models are 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 [1] to obtain the theme word distributions.

Specifically, let $\theta_1, \dots, \theta_k$ 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 (based on word generation).

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

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 $\Lambda = \{\theta_j, \pi_{d,j} | d \in C_i, 1 \leq j \leq k\}$, which can be estimated with an EM algorithm to maximize the log-likelihood of C_i :

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

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

2.3 Evolutionary Transition Discovery

To discover evolutionary transitions between any two theme sects, we use the Kullback-Leibler divergence [2] 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 sects 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 sects γ_1 and γ_2 where $s(\gamma_1) < s(\gamma_2)$, we compute $D(\theta_2|\theta_1)$. If $D(\theta_2|\theta_1)$ is above a threshold ξ , we will infer that γ_1 evolves into γ_2 . The threshold ξ allows a user to flexibly control the strength of the extracted themes.

2.4 Analysis of Theme Life Cycles

The theme evolution graph discussed above can model ETPs in a microcosmic view, which extracts theme sects within each time intervals and structures them. In a macroscopical view, we are interested in modeling the global evolutionary patterns of themes over the whole text stream and in analyzing the life cycles of specific themes. In this section, we present a method based on Hidden Markov Models (HMMs) [6] to model the shifting among trans-collection themes and analyze their life cycles.

Formally, an HMM consists of a set of hidden states $S = \{s_1, \dots, s_n\}$, a set of observable output symbols $O = \{o_1, \dots, o_m\}$, 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 use an HMM to model the stochastic process of theme shifts in our text stream, we assume that the collection, which is represented as a long sequence of words, is stochastically generated from a sequence of unobservable theme models, where the shifts between themes are represented as the transitions between the states of an HMM that correspond to the extracted theme models. Such an HMM is constructed in the following way. We first extract k trans-collection themes from the collection. The language models corresponding to each theme, as well as the background model, are used as the hidden states to construct a fully connected HMM. The entire vocabulary V is taken as the output symbol set. The output probability distribution of each state is set to the multinomial distribution of words given by the corresponding theme language 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 [6].

Once the initial state probabilities and transition probabilities are estimated, the Viterbi algorithm [6] can be used to decode the text stream to obtain the most likely state sequence, i.e., the most likely sequence of theme shifts.

Once the whole stream is decoded with the labels of themes, we can use a fix-size sliding window of time to measure the strength of each theme at a time point ¹. The strength of theme i at time t is computed as:

$$Strength(i, t) = \frac{1}{W} \sum_{w \in [t - \frac{W}{2}, t + \frac{W}{2}]} \delta(w, i)$$

The normalized strength of theme i at t is computed as:

$$NStrength(i, t) = \frac{1}{\sum_{w \in V} c(w, W)} \sum_{w \in [t - \frac{W}{2}, t + \frac{W}{2}]} \delta(w, i)$$

where $\delta(w, i) = 1$ if w is labeled as theme i ; otherwise $\delta(w, i) = 0$. W is the size of the sliding window in terms of time. *Strength* measures the absolute amount of text information about a given theme, while *Normalized Strength* measures the relative strength of this theme to other themes. The life cycle of each theme can then modeled as the variation of the theme strength or normalized strength over time.

¹The use of a sliding window also solves the “report delay” problem in the news domain.

3 Experiments

We evaluated the proposed ETP discovery methods on three data sets in two different domains - news articles and scientific literature.

3.1 Data Preparation

Two data sets are constructed to test our approaches. 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. Some documents which were not recognizable by the pdf2text software in Linux were excluded. The basic statistics of Tsunami news data and KDD Abstract data 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
SIGIR Full Text	1170	3439.94	1978-2004

Table 2. Basic information of data sets

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

3.2 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 2 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 sect. We see that most of these themes suggest meaningful subtopics in the context of the Asia tsunami event.

With these theme sects, we use the KL-divergence to identify evolutionary transitions. Figure 1 shows a theme evolution graph discovered from Asia Tsunami data when the threshold for evolution distance is set to $\xi = 12$. From Figure 1, we can see several interesting evolution threads which are annotated with symbols. By extracting these evolutionary threads and organizing them with temporal order, a summarization

²<http://www.acm.org/dl>

	Theme 1	Theme 2	Theme 3	Theme 4	Theme 5	Theme 6
11:	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
12:	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
13:	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
14:	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	Hafin 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
15:	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 sects extracted from Asia Tsunami data

of report on Tsunami events is shown in Figure 2.

The second experiment aims to model the patterns of life cycles of trans-collection themes. In this experiment, we analyze the life cycles of trans-collection themes in 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.

In Figure 3 we show the life cycles of the trans-collection themes in CNN by plotting the theme strength values over time. In Figure 3, we show the theme strength values of the global significant themes over time in Xinghua News.

3.3 Experiments on Scientific Literature

As in the news data, we also analyzed the life cycles of trans-collection themes in KDD Abstracts. Seven dominating trans-collection themes and their life cycles are presented in Figure 4. 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 4, we see that the normalized strength of the theme that suggests a traditional application topic of data mining which corresponds to marketing, business and customer analysis is decreasing all over six years. 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.

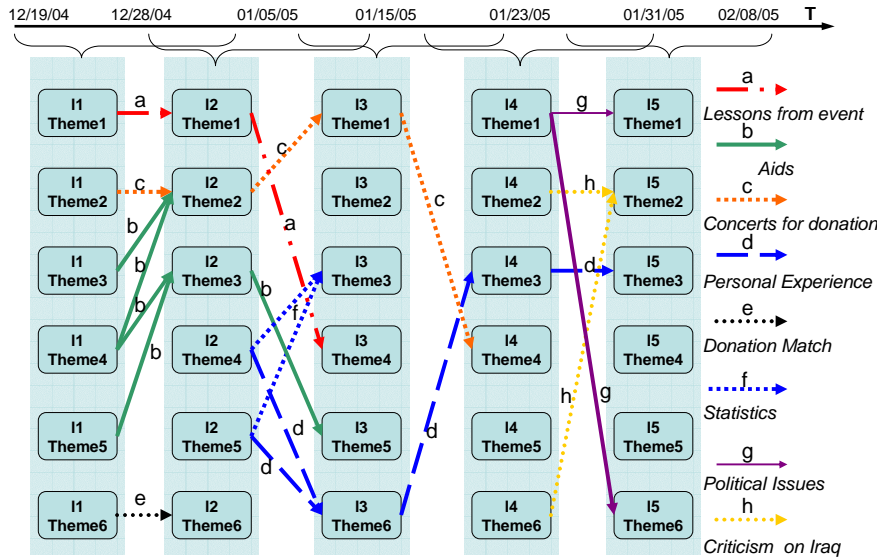


Figure 1. Theme evolution graph for Asia tsunami

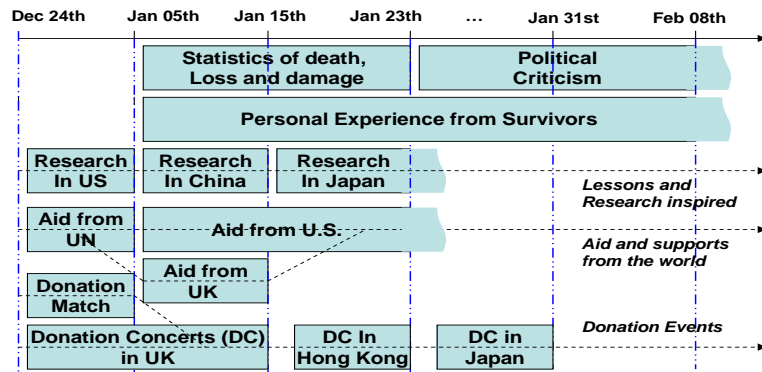


Figure 2. Summarization of Tsunami Reports based on Evolutionary Theme Threads

There are also themes, such as classification and clustering (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 all cover clustering, we may also infer that clustering is another major dominating theme in KDD publications.

Comparable patterns are revealed from SIGIR data set, which is shown in Figure 4.

4 Discussion

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 digest 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

Source	Theme 1	Theme 2	Theme 3	Theme 4	Theme 5
CNN	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
	Ocean 0.0073	beach 0.0056	hospital 0.0072	said 0.0054	million 0.0135
	Indian 0.0064	water 0.0051	baby 0.0064	military 0.0044	relief 0.0134
	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
XINHUA	Thailand 0.0104	Aceh 0.0219	Chinese 0.0391	dollars 0.0226	system 0.0314
	Thai 0.0096	province 0.0111	China 0.0391	million 0.0204	warning 0.0272
	missing 0.0079	Indonesian 0.0075	yuan 0.0180	US 0.0178	early 0.0172
	victims 0.0054	tidal 0.0055	countries 0.0098	aid 0.0118	meeting 0.0159
	Philippine 0.0040	waves 0.0047	Beijing 0.0089	United 0.0108	Ocean 0.0121
	confirmed 0.0040	killed 0.0045	travel 0.0061	countries 0.0106	small 0.0096
	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

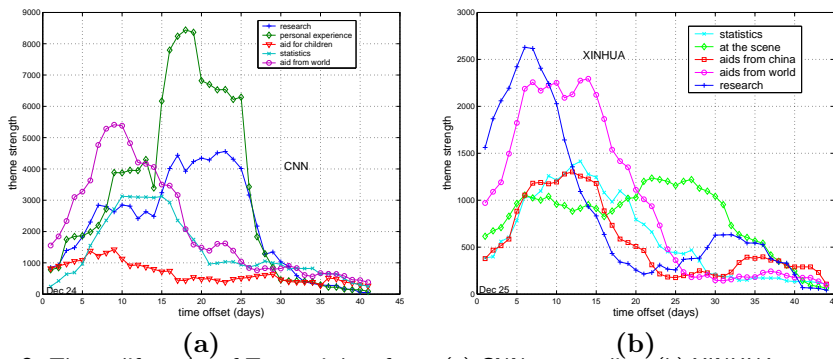


Figure 3. Theme life cycles of Tsunami data from: (a) CNN news online; (b) XINHUA news agency

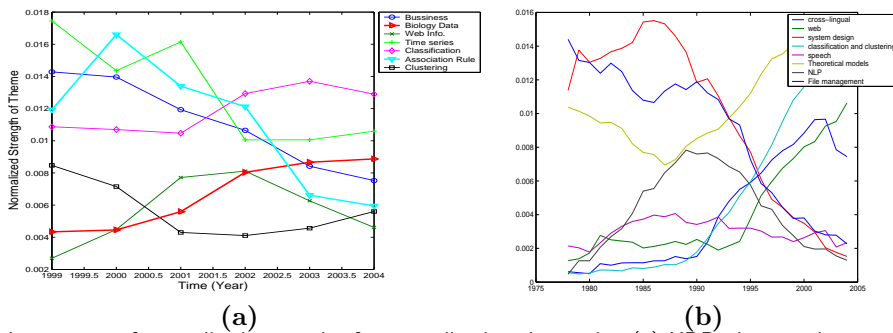


Figure 4. Life cycle patterns of normalized strength of trans-collection themes in: (a) KDD abstract data set; (b) SIGIR full text data set

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.

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 to study.

References

- [1] N. M. L. A. P. Dempster and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statist. Soc. B*, 39:1–38, 1977.
- [2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.
- [3] R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In *KDD*, pages 112–117, 1995.
- [4] M. A. Hearst. Untangling text data mining. In *Proceedings of the 37th conference on Association for Computational Linguistics (ACL 1999)*, pages 3–10, 1999.
- [5] 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.
- [6] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–285, Feb. 1989.
- [7] 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.