Generating Breakpoint-based Timeline Overview for News Topic Retrospection

Po Hu*, Minlie Huang*, Peng Xu†, Weichang Li†, Adam K. Usadi† and Xiaoyan Zhu*
* *State Key Laboratory on Intelligent Technology and Systems*
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, China
Email: chimera.linux@gmail.com, {aihuang, zxy-dcs}@tsinghua.edu.cn
†*ExxonMobil Research and Engineering Company, New Jersey, U.S.A*
Email: {peng.xu, weichang.li, adam.k.usadi}@exxonmobil.com

***Abstract*—Though news readers can easily access a large number of news articles from the Internet, they can be overwhelmed by the quantity of information available, making it hard to get a concise, global picture of a news topic. In this paper we propose a novel method to address this problem. Given a set of articles for a given news topic, the proposed method models theme variation through time and identifies the breakpoints, which are time points when decisive changes occur. For each breakpoint, a brief summary is automatically constructed based on articles associated with the particular time point. Summaries are then ordered chronologically to form a timeline overview of the news topic. In this fashion, readers can easily track various news topics efficiently. We have conducted experiments on 15 popular topics in 2010. Empirical experiments show the effectiveness of our approach and its advantages over other approaches.**

***Keywords*-Data Mining; Text Mining; News Topic Retrospection; Breakpoint;**

I. Introduction

The rapid development of the Internet has greatly influenced the way people digest information. With the prevalence of online news media, more and more news readers prefer reading news online. In 2008, for the first time in the U.S., the number of online news readers surpassed the number of newspaper readers [1]. This trend is expected to continue as more news sites emerge and investments in online news media grow[2].

The booming of the online news industry has made large numbers of news resources available on the Web. However, even with popular news search engines (e.g., Google News, Yahoo! News) and news-reading assistant tools (e.g., Google Alerts[1], Google News Timeline[2]), online news readers still face an information overloading problem. A reader may be overwhelmed by the huge number of news articles, therefore cannot identify the important dates or key phases of a news topic, particularly for those that are continuously developing. Table I shows such a situation for the topic "2010 Haiti Earthquake".

[1]http://www.google.com/alerts
[2]http://newstimeline.googlelabs.com

News Access Channels	Google News	Google Alerts	Google News Timeline
Type	Search Engine	E-mail Update	Web Application
Information Available	67,900 news articles	28 news updates per week	A graphical timeline with 10 news/month

Table I
Information overloading shown for the topic "2010 Haiti Earthquake".

From Table I, we see clearly that neither the conventional search engine (i.e., Google News) nor the news update tool (i.e., Google Alerts) is suitable for reviewing a news topic efficiently. The former retrieves too many articles, while the latter floods a user's mailbox with every new topic entry. Google News Timeline presents retrieved results in a better way, as news articles are organized chronologically, and a small number of selected news items are shown for each month. However, Google News Timeline, just like conventional search engines, selects news articles according to document relevance but not importance. Moreover, they do not detect key phases of a topic accurately, making it hard to provide a reliable chronological picture of a topic.

In this paper, we study the problem of automatic generation of breakpoint-based timeline overview for news topic retrospection. The timeline overview can provide readers with a clear and chronological view of the topic, and relieve them from reading news articles one by one. A *breakpoint* is defined as a date when decisive changes occur in the development of a topic. Breakpoints can manifest key phases of a news topic and depict the outline of the whole story. We develop a novel method that identifies breakpoints by modeling topic activeness and theme variation. A brief summary for each breakpoint is automatically constructed based on articles associated with the breakpoint. All summaries are ordered chronologically to form the timeline overview.

The rest of the paper is organized as follows. We briefly

survey related work in section 2. In section 3, we formally define the problem and present our methodology. Details of the approach are presented in sections 4, 5 and 6. Empirical experiments and analysis are shown in section 7. We conclude our work in section 8.

II. Related Work

Our work is related to several lines of research in the literature such as topic detection and tracking [3], [4], [5], news event detection and tracking [6], [7], temporal text mining [8], [9], [10], temporal news summarization [11], [12], [13], [14], [15], and Hidden Markov Model (HMM)-incorporated topic modeling [16], [17].

Lin and Liang [13] propose a storyline-based summarization for news topic retrospection. They use a self-organizing map to detect events in a news topic and then remove irrelevant events via event similarity comparison. The remaining events are organized sequentially as the storyline of the topic. The summaries of the events are used to form the summary of the whole topic. Our work is different from Lin and Liang's in that our method relies on breakpoint detection but not event detection. Although breakpoints are associated with key events, breakpoints can depict a topic's key phases in a finer and more flexible manner than key events.

Allan et al. [11] developed a temporal news summarization method, which is also based on event detection. In this method, content of the summary changes dynamically as new events are identified through the course of the topic. Although our work is also a kind of temporal news summarization, our goal is different from [11]. We aim at news topic retrospection, where timeline overview is generated without dynamic news update.

Our work is also related to topic detection and tracking (TDT) and temporal text mining (TTM). Mei and Zhai [9] study theme evolutions in a given news topic using a probabilistic mixture model and a hidden Markov model. Wang et al. [4] and Wang et al. [5] work on topic tracking from multiple news streams. Their methods extract meaningful topics from multi-source news collections, and track different topics as they evolve from one to another along the timeline. These methods focus on the evolving relationships among different topics, while ours focuses on a specific news topic, and detect the most important moments of the topic in its life cycle.

The research closest to ours is the breakpoints identification work by Akcora et al. [18]. They propose a simple set space model to detect changes in lexical patterns in tweets. Note that tweets are much shorter and less topic coherent than news articles. As a result, our approach is more sophisticated in that we utilize topic model and hidden Markov model to detect theme variation. We also propose a timeline overview based on the discovered breakpoints, while their method does not.

Finally, our breakpoint detection approach is relevant to HMM-incorporated topic modeling. Griffiths et al. [16] combine a hidden Markov model (HMM) with a latent Dirichlet allocation model (LDA) to construct the LDA-HMM model. LDA-HMM considers both syntactic relationships and semantic topics, and is used to identify function words and content words in a document. Gruber et al. [17] propose a hidden topic Markov model (HTMM), which incorporates a HMM into a topic model (LDA). When analyzing a document, the HTMM model assumes that all words in the same sentence belong to the same topic, and successive sentences are more likely to have the same topics. Thus the topic assignment in HTMM is more coherent compared to traditional "bag-of-words" assignments. In this paper, we also utilize topic model and HMM to model topic transition in news streams. Similar to LDA-HMM and HTMM, our approach assumes that topic transitions within a news stream form a Markov chain. However, our approach is much simpler than the above two models in that transition probabilities of the topic Markov chain are independent of the topic sampling process in the topic model.

III. Problem Formulation and Methodology

Suppose a user wants to review a news topic Q, which has a stream of relevant news articles $\{D_t|t = 1, ..., T\}$, where D_t is a document collection containing news articles about Q published at time point t. In this paper, each t lasts for a day.

We define the breakpoints of Q as follows.

Definition 1 (BREAKPOINT): A breakpoint is a specific time point in $\{1, ..., T\}$, when a decisive change or significant development occurs in the topic progress. The change can be signified by the beginning, breaking, or ending of some key events.

Take the news topic "Chile Miner Rescue" for example. The breakpoints of the topic are the dates when the accident occurred, when the drilling of the escape route began, and when the trapped miners were successfully rescued, etc.

The breakpoint-based timeline overview of Q is defined as follows.

Definition 2 (OVERVIEW): A timeline overview[3] is an ordered sequence of $\mathcal{M}$ summaries with respect to $\mathcal{M}$ breakpoints of Q. Each summary is a brief review of events happening at the corresponding breakpoint.

News timeliness is the foundation of our methodology. The intuition is that news articles posted near a breakpoint may report on the triggers, details, and consequences of the critical change, thus the themes of news articles before a breakpoint may be quite different from those of articles at the breakpoint. Therefore, we can make the following assumption:

[3]Examples of timeline overview are shown in Figures 3 - 6 on page 10.

A breakpoint can be detected by analyzing theme variation before and after a point.

In line with this assumption, we can discover a breakpoint using news articles that are chronologically near that point, instead of using the entire news stream. We therefore slice the whole period of a topic into N equal-length time intervals, each of which covers certain time points (intuitively, each interval contains about seven points corresponding to the weekly business cycle). Adjacent intervals have overlapping time points to cope with the time delay among news articles from different agencies. The entire news collection $\{D_1, ..., D_T\}$ is also divided into N subcollections according to the partition of time points.

Topic activeness plays a key role in breakpoint detection. The activeness of a news topic varies in its life cycle. There are different topic activeness patterns, depending on the nature of the topic. Some burst in the beginning and gradually fade out, while others may have a humble start but develop dramatically thereafter. Breakpoints can appear intensively in an active interval, but has little chance to appear in an inactive interval. Therefore, we may discard those topic inactive intervals (referred as *null intervals*) before mining the breakpoints from each interval in the topic's life cycle. Removing null intervals can effectively avoid catching false breakpoints, thus improving the accuracy of breakpoint detection.

Our approach consists of three steps.

Step 1: Analyze each interval's topic activeness using the *Topic-Activeness HMM model*, and discard inactive intervals.

Step 2: Identify breakpoints by detecting theme variation in each time interval, through a topic mixture model and the *Theme-Transition HMM model*.

Step 3: Generate a summary for each breakpoint by selecting representative sentences, and construct a timeline overview according to the summaries of the discovered breakpoints.

IV. Topic Activeness Modeling

A topic's activeness in an interval correlates with two factors: 1) the news quantity in the interval, 2) the new information the interval provides compared to the preceding interval.

On one hand, an interval with few articles indicates that there are no valuable stories in this period, thus the topic activeness is low and no breakpoint exists. On the other hand, if an interval provides little new information, i.e. the content of the news articles in the interval is similar to those in the preceding one, indicating the reflected progress of the topic is very slow, or even temporarily stops during this interval. Therefore the topic is inactive in terms of topic development, and such an interval may not contain any breakpoint.

We propose a *Topic-Activeness HMM model* to analyze topic activeness for each time interval. The Topic-Activeness HMM model is a first-order fully-connected hidden Markov model [19]. Suppose there are M levels of topic activeness, and each level is represented by a hidden state in ascending order $s_i \in \{s_1, ..., s_M\}$ with respect to topic activeness. The HMM's observation symbols are the numbers of news articles published in each interval. We set the emission probability distributions of the hidden states as Poisson distributions

$$B_{i,k} = \frac{{\lambda_i}^k e^{-\lambda_i}}{k!} \tag{1}$$

where $B_{i,k}$ is the emission probability for state s_i to generate k news articles in an interval and λ_i is the expected articles count of the interval, corresponding to state s_i. For clarity, we define the *size* of an interval as number of news articles in the interval. To determine the value of λ_i, traditional methods (as used in [20]) first divide the whole range of all interval sizes into M equal-sized bins, then choose the mean interval size in the ith bin as the value for λ_i. This method, denoted as *equal-size binning*, performs poorly when the distribution of interval sizes is highly uneven (as is often the case). Figure 1 illustrates such a situation by an extremely uneven distribution of interval sizes.

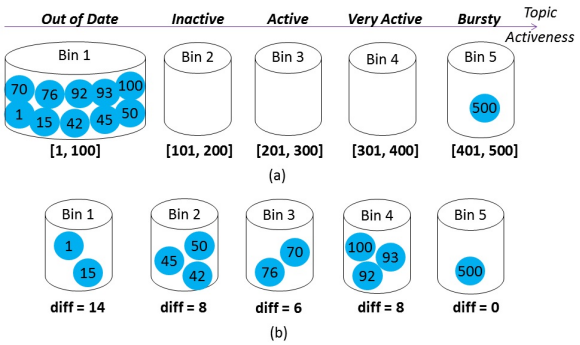

Figure 1. An illustration of interval size division: (a) equal-size binning. (b) dynamic binning (each circle represents a time interval, the number in the circle is the interval size).

In this example, there are 11 time intervals with number of news articles in ascending order: {1, 15, 42, 45, 50, 70, 76, 92, 93, 100, 500}, and there are 5 activeness levels. As shown in Figure 1 (a), equal-size binning will divide the whole range of news articles counts (i.e. [1, 500]) into 5 equal-size bins. All intervals except the one valued 500 are allocated into the first bin.

To adequately handle the uneven distribution of news articles, and dynamically set the thresholds for various topics, we propose a *dynamic binning* method to replace the equal-size binning method. In dynamic binning, time intervals (with their articles counts) are grouped into M clusters (bins); each cluster corresponds to a topic activeness level. The *dispersion* of a cluster is defined as the difference between the largest and the smallest interval size in the

cluster. The goal of dynamic binning is to minimize the largest dispersion of all clusters. Figure 1 (b) shows the optimal solution of dynamic binning. The dispersion of each cluster is 14, 8, 6, 8, 0, respectively. This balanced dispersion enables a more reliable estimation of the values of λ_i as explained late in this section.

Dynamic binning can be solved efficiently using dynamic programming and the best solution is always available. The complexity of dynamic binning is $O(MC^2)$, where M is the number of clusters and C is the number of time intervals.

Based on dynamic binning, we determine the value of λ_i with the following procedures. As mentioned before, the topic's activeness correlates with number of news articles and new information available in an interval. So firstly, we calculate the new information in each interval compared to the previous one. This is achieved by measuring the KL-divergence of word frequency distributions. Formally, the new information an interval Γ_i provides is defined as

$$D_{KL}(\Gamma_i||\Gamma_{i-1}) = \sum_{w \in W} p(w|\Gamma_i) \log \frac{p(w|\Gamma_i)}{p(w|\Gamma_{i-1})} \quad (2)$$

where W contains all the stemmed words in both intervals Γ_{i-1} and Γ_i, $\{p(w|\Gamma_i)\}_{w \in W}$ is the word frequency distribution of Γ_i. The first interval (i.e., i = 1) is compared to an empty interval, resulting in a very high KL-divergence for the first interval. This result reflects the fact that the first interval contains the starting point of the topic, which is a very important breakpoint.

Secondly, we combine the value of news quantity and new information for each interval by multiplying the number of articles and KL-divergence. Then we normalize the combined value of each interval to the range of interval sizes as

$$Value_{Nor} = (\frac{MAX - MIN}{MAX^{'} - MIN^{'}})(Value - MIN^{'}) + MIN \quad (3)$$

where MAX and MIN are the maximum and minimum of the interval size, $MAX^{'}$ and $MIN^{'}$ are the maximum and minimum of (articles count×KL-divergence). $Value_{Nor}$ is the normalized value of the original value $Value$. $Value_{Nor}$ can be interpreted as the *fixed size* of an interval, considering the new information available in the interval.

Finally, we sort the fixed sizes of all intervals in ascending order, and group the values into M bins using dynamic binning. We set the value of λ_i as the mean size of the intervals in the ith bin, and interpret it as the *expected articles count* of an interval corresponding to state s_i.

The parameters in Topic-Activeness HMM are estimated using Baum-Welch algorithm [19]; topic activeness for each interval is obtained using Viterbi algorithm [19]. The result is a sequence of topic activeness levels, each of which corresponds to an interval in the topic's life cycle. Those intervals labeled as low active levels are considered as null intervals, and are eliminated for breakpoint detection.

V. Breakpoint Detection

In general themes of news articles posted near a breakpoint change noticeably before and after that date. Thus we discover breakpoints within a l-day interval $\Gamma = (t_i, t_{i+1}, ..., t_{i+l-1})$ by two steps: theme extraction and theme variation detection.

A. Theme Extraction

We use a generative probabilistic mixture model [9] to extract themes of the news collection in Γ. Suppose there are K themes $\{z_1, ..., z_K\}$ and a background theme z_B in the document stream $\mathcal{D} = (D_{t_i}, D_{t_{i+1}}, ..., D_{t_{i+l-1}})$ within Γ. A theme z_k[4] is a probabilistic distribution of words in the word set $\mathcal{V}$ of $\mathcal{D}$; that is, z_k governs the multinomial word distribution $\{p(w|z_k)\}_{w \in \mathcal{V}}$. A word w in a document $d \in \mathcal{D}$ is sampled according to the following probability:

$$p(w|d) = \omega_B p(w|z_B) + (1 - \omega_B) \sum_{k=1}^{K} p(z_k|d) p(w|z_k) \quad (4)$$

where ω_B is the weight for z_B, and is selected ad hoc. The background theme z_B, by definition, is formed with high-frequent but low-informative words (i.e., domain stopwords). As a result, the K themes we extract are more discriminative and meaningful due to the introduction of z_B. The occurrence of a word w given a background theme can be estimated as $p(w|z_B) = \frac{\sum_{d \in \mathcal{D}} c(w,d)}{\sum_{w^{'} \in \mathcal{V}} \sum_{d \in \mathcal{D}} c(w^{'},d)}$, where $c(w, d)$ is the count of word w in document d.

By maximizing the following log likelihood, the model's parameters, $\{p(z_k|d)\}$ and $\{p(w|z_k)\}$, can be obtained by using EM algorithm [21]:

$$\log p(\mathcal{D}) = \sum_{w \in \mathcal{V}} \sum_{d \in \mathcal{D}} c(w, d) \log p(w|d) \quad (5)$$

B. Theme Variation Detection

We treat each document as a word stream, and the word streams of all documents in D_{t_j} is concatenated as a word sequence without considering the order of the documents. Hence, the word sequence of $\mathcal{D}$ in interval Γ is formed by sequentially concatenating all word streams at all time points in Γ. Each word in the word sequence of $\mathcal{D}$ is generated by some theme $z \in \{z_1, ..., z_K, z_B\}$. Our task is to find the most probable theme sequence that generates this word sequence, which is solved by the *Theme-Transition HMM model*.

The Theme-Transition HMM model is a first-order fully-connected hidden Markov model [19] that is used to determine the theme shift sequence in $\mathcal{D}$. The HMM model has $K + 1$ hidden states corresponding to the $K + 1$ themes $(z_1, ..., z_K, z_B)$ we have extracted. The observation sequence

[4]In this section, k represents the index of a theme, not the articles count as used in Section IV.

of HMM is the word sequence of $\mathcal{D}$, and the emission probability distribution is $\{p(w|z)\}$. The initial state probabilities and transition probabilities are estimated using Baum-Welch algorithm [19]. We then obtain the most probable theme shift sequence that generates the word sequence of $\mathcal{D}$ using Viterbi algorithm [19]. Note that the same word within a news article may belong to different topics depending on the context. Figure 2 is an illustration of the Theme-Transition HMM model.

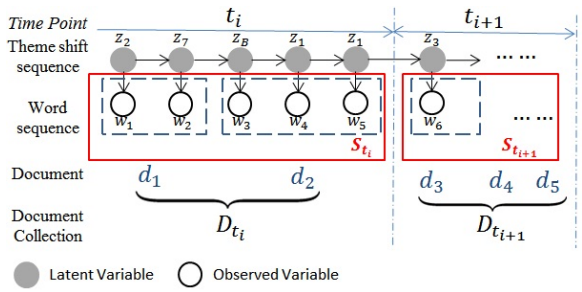

Figure 2. An illustration of theme shift modeling using the Theme-Transition HMM model.

We define the strength of z at time point t, $\sigma(z,t)$, as the number of words generated by z in the word sequence of D_t, normalized by the total number of words in D_t. Each time point t is associated with a theme strength distribution $\{\sigma(z_1,t),...,\sigma(z_K,t),\sigma(z_B,t)\}$. The stronger a theme's strength is, the more dominant the theme is at that time point.

Once the distributions are obtained, we can detect breakpoints in the interval by measuring the variation of theme strength distributions across time points. We use a distance metric, the square root of the Jensen-Shannon divergence [22], to measure the theme variation $TV_{(t,t+1)}$ of two neighboring time points t and $t+1$, as follows:

$$\mathrm{TV}_{(t,t+1)} = \sqrt{\begin{array}{l} \frac{1}{2}\sum_{k=1}^{K+1}\sigma(z_k,t)\log\frac{\sigma(z_k,t)}{m_{z_k}} + \\ \frac{1}{2}\sum_{k=1}^{K+1}\sigma(z_k,t+1)\log\frac{\sigma(z_k,t+1)}{m_{z_k}} \end{array}} \tag{6}$$

where $m_{z_k} = \frac{1}{2}[\sigma(z_k,t)+\sigma(z_k,t+1)]$, and z_{K+1} is the background theme z_B.

Time point t is considered to be a breakpoint if the following criterion is satisfied:

$$\begin{cases} TV_{(t-1,t)} > TV_{(t-2,t-1)} \\ TV_{(t-1,t)} > TV_{(t,t+1)} \end{cases} \tag{7}$$

This means that there is a noticeable theme change at breakpoint t comparing with the immediate preceding and subsequent time points.

Note that the above criterion does not consider the first two time points and the last point of the interval. For these points, we compare the maximum theme strength of the test point to the averaged maximum theme strength at the time points in the interval. If the point's theme strength is larger than the averaged strength, we consider the time point as a breakpoint.

VI. Timeline Overview Generation

A timeline overview is generated by selecting key sentences to form a summary for each breakpoint, and chronologically ordering the summaries for all breakpoints of a topic.

We note that the title and the first sentence of the main body from a news article briefly summarize the content of the article. Therefore, these two sentences are extracted from each news article published at breakpoint t_b, to construct the candidate sentences for t_b's summary. The sentences are then clustered and a representative sentence from each cluster is selected to form t_b's summary. The representative sentence is selected as follows. First, a topic signature word set TW_{t_b} is generated by extracting 10 words with highest $p(w|z)$ for the top two high-strength themes at t_b. Second, for each sentence, a word set is formed by extracting the informative words (i.e., noun, verb, adj., adv.) from the sentence. Third, each sentence is ranked by measuring its similarity between its word set and the topic signature word set. The higher the similarity, the higher the rank. Jaccard Similarity[5] is used as the similarity metric.

Algorithm 1 shows the pseudo code of our overview generation method.

VII. Empirical Experiments

A. Dataset and Gold Standard

We build a news dataset from five well known news agencies: ABC, BBC, FOX, Reuters and USAToday. We also select 15 news topics that had been the hottest spots in 2010. These topics cover a wide range of issues, including natural disasters, politics, finance, conflicts, and accidents. For each news topic, we download relevant news articles from the five news agencies using keyword search embedded in their websites and restrict the news articles to be within a certain period of time.

The first step to constructing a gold standard is to generate a list of breakpoints for each topic. For each topic, we collect human-written timeline overviews from five authoritative online media, including four news agencies (CNN, New York Times, BBC, and Reuters) and Wikipedia. Each human written overview contains a list of summaries where each summary corresponds to a time point. Since each human-written overview has its own focus on the topic, and is written with different goals in mind, we need to determine the agreement between them. We select the time points that

[5]Defined as $Sim_{Jac}(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$, A/B are two sets.

Algorithm 1 Breakpoint-based timeline overview

Input: set of breakpoints P; topic signature words TW_{t_b} for each breakpoint t_b; maximum summary length L;
Output: a timeline overview OVERVIEW;

```
OVERVIEW ⇐ Φ;
for each breakpoint t_b ∈ P do
    Sum_{t_b} ⇐ Φ;
    for each news article in t_b do
        Extract the news headline and first sentence of the main body into a candidate sentence set S_{t_b};
    end for
    for each sentence s_j ∈ S_{t_b} do
        Extract the informative words (n./v./adj./adv.) of s_j, and put them into a word set I_{s_j};
    end for
    Use k-means to cluster the sentences in S_{t_b}, each of which is represented by I_s respectively, with Jaccard Similarity as the distance metric;
    for each cluster c do
        Select sentence s in c with the largest |I_s|, where |I_s| is the cardinality of I_s, as the representative sentence of cluster c;
    end for
    Iteratively put representative sentence s with highest Sim_Jac(I_s, TW_{t_b}) into Sum_{t_b}, while the length of Sum_{t_b} < L;
    Put Sum_{t_b} into OVERVIEW;
end for
```

are presented in at least three human written overviews as the gold standard of the topic's breakpoints. If there is little agreement (which occurred in 5 topics from our dataset) only the dates in Wikipedia overview are chosen as breakpoints. Once a breakpoint is selected, all human written summaries for that date are collected to form the breakpoint's summary.

Table II is a brief description of our dataset. The ratio of breakpoints in the gold standard to the total number of time points of a topic ($\frac{\#Breakpoints}{\#TimePoints}$) ranges from 12% to 22%. This shows that breakpoints are just a small proportion of the total time points for a news topic, which is consistent with our common sense.

News Topic	Brief Introduction	#News/#P TimeSpan
Deepwater Horizon	Oil spill in the Gulf of Mexico as a result of the explosion of oil drilling rig Deepwater Horizon.	2272/19 6 months
Health Care Reform	Debates and challenges regarding 2010 health care reform in U.S.	964/17 6 months
Samho Dream	Rescue of supertanker Samho Dream from Somali pirates.	448/29 6 months
Haiti Earthquake	A magnitude 7.0Mw earthquake hit Haiti killed 316,000 people.	2072/20 4 months
EU Finance Crisis	Severe financial crisis outbroke in several EU members.	1426/15 4 months
Iceland Volcano Eruption	Severe air travel disruption due to the volcanic ash from a series of Iceland volcano eruptions.	1384/19 4 months
Greek Riots	Nationwide riots across Greece due to government debt crisis.	424/24 4 months
Pakistan Flood	Heavy monsoon rains caused floods affected 20 million people.	761/16 3 months
Thailand Protest	Violent protests held by Red-Shirts against Prime Minister.	642/18 3 months
French Strikes	Series of general strikes across France against pension reform.	568/9 3 months
Chile Miner	The rescue of 33 Chilean miners trapped underground for 70 days.	523/11 3 months
Senkaku Collision	Conflicts over the Diaoyu or Senkaku boat collision incident.	245/11 3 months
Russia Wildfires	Hundreds of wildfires broke out in Russia due to extreme heat.	216/9 3 months
Poland Plane Crash	Polish plane crashed in Russia killed president Lech Kaczynski.	269/12 2 months
Anna Chapman	The biggest spy swap since Cold War between Russia and U.S.	217/7 2 months

Table II
A BRIEF DESCRIPTION OF OUR DATASET. #P IS THE NUMBER OF BREAKPOINTS IN THE GOLD STANDARD.

B. Experiment Results and Analysis

We evaluate our approach by three sets of experiments: 1) null interval identification, 2) breakpoint detection, and 3) timeline overview generation. Each set of experiments corresponds to a key step in our breakpoint detection approach. The Wilcoxon signed-rank test [23] is used for statistical significance test in our experiments.

First, we investigate the performance of Topic-Activeness HMM in null interval identification. We determine the number of hidden states, M, in the model and the inactive levels that signal null intervals. As discussed in [20], the more hidden states, the finer the topic activeness can be depicted. However, more hidden states will increase the complexity of HMM training, and the performance improvement may be negligible once there are enough hidden states. Thus, we adopt the optimal parameter settings reported in [20] for the Topic-Activeness HMM model, i.e. there are 5 hidden states, and states s_1 and s_2 are used to signal null intervals.

We also design three baselines to compare to our model. Baseline 1 (Article Count) only use number of news articles in each interval to determine the null intervals. Dynamic binning is used to divide the articles counts into 5 bins, and the top 2 bins with the smallest articles count are considered to be null intervals. Baseline 2 (KL) and Baseline 3 (Count×KL) are similar to Baseline 1; the only difference is that we use KL-divergence and (articles count×KL-divergence) to replace the articles count used in Baseline 1, respectively. We also investigate the ratio of null intervals to all intervals for each topic. Table III shows the experiment results for null interval identification.

From Table III we see clearly that the ratio of null intervals varies greatly from one topic to another (from 14.3% to 54.5%). On average, 35.8% intervals of a topic are null intervals, indicating that null interval identification is essential to the breakpoint detection approach. Topic-Activeness HMM outperforms all three baselines in precision (30.0% better than the second best one) and F_1

Topic	Null Interval	Metric	Articles Count	KL	Count ×KL	Ours	Topic	Null Interval	Metric	Articles Count	KL	Count ×KL	Ours
Deepwater Horizon	48.0%	P	53.8	55.6	57.1	**62.5**	Health Care Reform	54.5%	P	81.8	60.0	64.7	**83.3**
		R	58.3	**83.3**	66.7	**83.3**			R	75.0	50.0	**91.7**	83.3
		F_1	56.0	66.7	61.5	**71.4**			F_1	78.3	54.5	75.9	**83.3**
Samho Dream	52.3%	P	53.6	**65.0**	52.8	50.0	Haiti Earthquake	23.1%	P	33.3	20.0	30.0	**66.7**
		R	65.2	56.5	**82.6**	78.3			R	**100**	66.7	**100**	66.7
		F_1	58.8	60.5	**64.4**	61.0			F_1	50.0	30.8	46.2	**66.7**
EU Finance Crisis	47.4%	P	57.1	54.5	57.1	**70.0**	Iceland Volcano Eruption	29.4%	P	**41.7**	36.4	38.5	**41.7**
		R	**88.9**	66.7	**88.9**	77.8			R	**100**	80.0	**100**	**100**
		F_1	69.6	60.0	69.6	**73.7**			F_1	**58.8**	50.0	55.6	**58.8**
Greek Riots	54.2%	P	70.6	57.1	61.9	**78.6**	Pakistan Flood	33.3%	P	33.3	33.3	44.4	**50.0**
		R	92.3	61.5	**100**	84.6			R	50.0	75.0	**100**	50.0
		F_1	80.0	59.3	76.5	**81.5**			F_1	40.0	46.2	**61.5**	50.0
Thailand Protest	14.3%	P	20.0	25.0	22.2	**40.0**	French Strikes	25.0%	P	25.0	25.0	25.0	**66.7**
		R	**100**	**100**	**100**	**100**			R	**66.7**	**66.7**	**66.7**	**66.7**
		F_1	33.3	40.0	36.4	**57.1**			F_1	36.4	36.4	36.4	**66.7**
Chile Miner	33.3%	P	50.0	50.0	44.4	**80.0**	Senkaku Collision	40.0%	P	60.0	40.0	50.0	**75.0**
		R	**100**	75.0	**100**	**100**			R	**75.0**	50.0	**50.0**	**75.0**
		F_1	66.7	60.0	61.5	**88.9**			F_1	66.7	44.4	50.0	**75.0**
Russia Wildfires	28.6%	P	50.0	50.0	50.0	**66.7**	Poland Plane Crash	33.3%	P	45.5	41.7	41.7	**62.5**
		R	**100**	**100**	**100**	**100**			R	**100**	**100**	**100**	**100**
		F_1	66.7	66.7	66.7	**80.0**			F_1	62.5	58.8	58.8	**76.9**
Anna Chapman	20.0%	P	**50.0**	33.3	33.3	**50.0**	**Average**	35.8%	P	48.4	43.1	44.9	**62.9**
		R	**100**	**100**	**100**	**100**			R	84.8	75.4	**89.8**	84.4
		F_1	**66.7**	50.0	50.0	**66.7**			F_1	59.4	52.3	58.1	**70.5**

Table III
EXPERIMENT RESULTS OF NULL INTERVAL IDENTIFICATION.

score (18.7% better than the second best one), and has a comparable recall to the best one (84.4% vs. 89.8%). The F_1 score of our approach is significantly better (p-value$<$0.5%) than the three baselines. This shows the effectiveness of Topic-Activeness HMM over simple statistical methods, and the benefit of smoothing in topic activeness transitions using the HMM model.

While our algorithm performs better than the baseline, it still discards a significant number of non-null intervals. Thus it is worth analyzing the performance of Topic-Activeness HMM in detail. For each topic, we calculate the ratio of the missed breakpoints due to the removal of false null intervals ("#P Missed"). We also check the ratio of the false null intervals that contain only one breakpoint to all the false null intervals ("1p False Interval"). Table V shows the performance analysis of the Topic-Activeness HMM model.

From Table V, we see that on average 89.2% of the misjudged null intervals contain only one breakpoint, indicating Topic-Activeness HMM is not good at distinguishing those intervals containing one breakpoint from the null intervals. Furthermore, on average, 15.0% breakpoints are missed due to the misjudging of false null intervals, showing that the overall performance of Topic-Activeness HMM is satisfactory. Finally, it is interesting to see that for some topics (such as "Health Care Reform" and "French Strikes"), the breakpoint missed is 0% even though Topic-Activeness HMM is suffered from null interval misjudging. This is due to the fact that adjacent time intervals are overlapped to deal with the news delay, and that the breakpoints in the false null intervals may also appear in their adjacent intervals. Thus those breakpoints may not be missed in breakpoint detection.

Topic	#P Missed	1p False Interval	Topic	#P Missed	1p False Interval
Deepwater Horizon	15.8%	100%	Health Reform	0%	100%
Samho Dream	31.0%	87.5%	Haiti Quake	15.0%	0%
EU Finance	26.7%	66.7%	Iceland Volcano	31.6%	85.7%
Greek Riots	8.33%	100%	Pakistan Flood	12.5%	100%
Thailand Protest	16.7%	100%	French Strikes	0%	100%
Chile Miner	18.2%	75%	Senkaku Collision	0%	100%
Russia Wildfires	11.1%	100%	Poland Crash	8.33%	100%
Anna Chapman	14.3%	100%	**Average**	**15.0%**	**89.2%**

Table V
PERFORMANCE ANALYSIS OF THE TOPIC-ACTIVENESS HMM MODEL.

Secondly, we compare our method to three baselines to evaluate the performance for breakpoint detection. Baseline 1 (Quant) identifies breakpoints by selecting the top 15% time points based on numbers of news articles. The intuition is that news agencies will publish intensively when a decisive change occurs, resulting in a dramatic increase in the number of news articles. Baseline 2 (Trend) is adapted from Akcora et al.[18], where breakpoints are detected via

Topic	Metric	Quant	Trend	Google	Ours	Topic	Metric	Quant	Trend	Google	Ours
Deepwater Horizon	P	25.0	27.5	32.6	**68.4**	Health Care Reform	P	**61.5**	28.9	37.1	52.6
	R	26.3	57.9	**78.9**	68.4		R	47.1	64.7	**76.5**	58.8
	F_1	25.6	37.3	46.1	**68.4**		F_1	53.3	40.0	50.0	**55.6**
Samho Dream	P	**45.0**	36.1	32.4	36.2	Haiti Earthquake	P	**77.8**	39.1	54.5	48.1
	R	31.0	44.8	**79.3**	58.6		R	35.0	45.0	60.0	**65.0**
	F_1	36.7	40.0	**46.0**	44.7		F_1	48.3	41.9	**57.1**	55.3
EU Finance Crisis	P	36.4	36.0	32.0	**50.0**	Iceland Volcano Eruption	P	**87.5**	57.1	58.3	76.9
	R	26.7	**60.0**	53.3	**60.0**		R	36.8	42.1	**73.7**	52.6
	F_1	30.8	45.0	40.0	**54.5**		F_1	51.9	48.5	**65.1**	62.5
Greek Riots	P	**78.6**	42.9	35.9	73.3	Pakistan Flood	P	50.0	42.1	47.1	**56.3**
	R	45.8	25.0	**58.3**	45.8		R	31.3	50.0	50.0	**56.3**
	F_1	**57.9**	31.6	44.4	56.4		F_1	38.5	45.7	48.5	**56.3**
Thailand Protest	P	**70.0**	44.0	47.8	57.9	French Strikes	P	25.0	27.8	41.2	**47.1**
	R	38.9	**61.1**	**61.1**	**61.1**		R	22.2	55.6	**88.9**	**88.9**
	F_1	50.0	51.2	53.6	**59.5**		F_1	23.5	37.0	56.3	**61.5**
Chile Miner	P	**71.4**	42.9	47.4	47.1	Senkaku Collision	P	**62.5**	53.9	38.1	56.3
	R	45.5	54.5	**81.8**	72.7		R	45.5	63.6	72.7	**81.8**
	F_1	55.6	48.0	**60.0**	57.1		F_1	52.6	58.3	50.0	**66.7**
Russia Wildfires	P	71.4	54.5	70.0	**85.7**	Poland Plane Crash	P	**87.5**	61.5	37.0	64.3
	R	55.6	66.7	**77.8**	66.7		R	58.3	66.7	**83.3**	75.0
	F_1	62.5	60.0	73.7	**75.0**		F_1	**70.0**	64.0	51.2	69.2
Anna Chapman	P	**83.3**	50.0	60.0	66.7	**Average**	P	**62.2**	43.0	44.8	59.1
	R	71.4	42.9	**85.7**	57.1		R	41.2	53.4	**72.1**	64.6
	F_1	**76.9**	46.2	70.6	61.5		F_1	48.9	46.3	54.2	**60.3**

Table IV
EXPERIMENT RESULTS OF BREAKPOINT DETECTION.

the variation of topic trend and lexical patterns over the entire news stream. Baseline 3 is Google News Timeline (Google), which selects around 10 news articles per month along the timeline, based on their relevance to the news topic. Although some other research mentioned in section II is relevant to our work, such as Lin and Liang [13], we don't compare their methods to ours due to the lack of experiment details and source codes.

The parameters used in our approach are set as follows. A time interval is 7 days long with 1 day overlap; the value is set by considering the publishing cycle of news articles. On average, a topic has less than 200 news articles in an interval, so we empirically set the topic number K to 7 for all datasets in the mixture model. ω_B controls the influence of the background theme z_B. As discussed in [9], the more noisy the text is, the more influential z_B should be. Since all words in the news are stemmed with stopwords removed (while [9] does not), the noise in the formatted news is relatively low. Thus the influence of z_B need not to be as strong as in [9] (where $\omega_B = 0.9$), and we empirically set ω_B to 0.2. The results of breakpoint detection are shown in Table IV.

Table IV shows our method outperforms Baseline 2 (Trend) substantially on all datasets. This is because we detect theme variation at a finer level of granularity, i.e. in a time interval, instead of the whole period of a topic. Since the content of the news collection within an interval is more semantically coherent than the content of the entire news stream, our method detects theme variation more accurately and sensitively than Baseline 2. Furthermore, our method gives the best averaged F_1 score (60.3%), achieving a good balance between precision and recall in breakpoint detection. Baseline 1 (Quant) has a poor averaged recall (41.2%) due to the "burst and diverse" feature [8] of news articles. When a news topic becomes active, news agencies will publish news articles on various aspects of the topic. Huge number of news articles diversify in content, rather than concentrate on the same key issues. So the burst in news quantity may not indicate that a critical change has happened. Our method also exceeds Baseline 3 (Google News Timeline) by 31.9% in the averaged precision (59.1% vs. 44.8%) and has a better F_1 score (60.3% vs. 54.2%). Since Google News Timeline selects the news on each month according to their relevance to the topic, the experiment result shows that our topic-based method is more accurate than Google News Timeline. The F_1 score of our approach is significantly better (p-value$<$0.5%) than the three baselines. Finally, we have noted that noise, which is introduced in the news crawling process, affects our algorithm's performance significantly for some topics (e.g., "Samho Dream"). This suggests that we need to improve our algorithm to be more robust to noise.

Thirdly, we evaluate our summarization algorithm against Google News Timeline (Google). For the baseline, we use news snippets generated by Google News Timeline to construct the summary of a breakpoint. In our algorithm, each breakpoint's summary has no more than 50 words, which is the maximum length for a news snippet in Google News Timeline. Since a candidate sentence has 10 words on average, a 50-word summary may contain 4-5 sentences. Therefore we set the cluster number in k-means to 7 to

generate 7 representative sentences to form the summary. ROUGE-1[24] is used to evaluate the summarization performance for all matched breakpoints, with stopwords removed. Table VI shows the overview evaluation results.

Topic	Google	Ours	Topic	Google	Ours
Deepwater Horizon	0.278	**0.390**	Health Reform	0.235	**0.333**
Samho Dream	0.283	**0.288**	Haiti Quake	0.144	**0.279**
EU Finance	0.070	**0.215**	Iceland Volcano	0.186	**0.320**
Greek Riots	0.115	**0.208**	Pakistan Flood	0.158	**0.359**
Thailand Protest	0.240	**0.400**	French Strikes	0.267	**0.336**
Chile Miner	0.153	**0.212**	Senkaku Collision	0.097	**0.408**
Russia Wildfires	0.216	**0.274**	Poland Crash	0.235	**0.322**
Anna Chapman	0.234	**0.312**	**Average**	0.194	**0.310**

Table VI
EXPERIMENT RESULTS OF TIMELINE OVERVIEW GENERATION.

Table VI shows that our summarization algorithm outperforms Google News Timeline substantially on all data sets, and surpasses the latter by 59.8% in terms of averaged ROUGE-1. This shows our method can better summarize the breakpoints of news topics than the popular online system, and the content of our timeline overviews is closer to the human-written overviews.

Excerpts of two news topic timeline overviews are shown in Figures 3 - 6 on page 10. We can see from Figure 3 that the overview of the topic "Chile Miner" successfully captures the key phases of the topic development, including the beginning of the accident, the turning points of the rescue work, and the successful release of all trapped miners. The 50 word summary for each breakpoint concisely describes the key events happened on the critical date. Thus the generated overview presents a clear, global picture of the prolonged rescue process to the reader. Figure 4 to 6 also show that our timeline overview can effectively catch and depict the critical changes of a news topic.

VIII. CONCLUSIONS

In this paper we address the problem of automatic generation of breakpoint-based timeline overview for news topic retrospection. Our approach detects breakpoints by modeling global topic activeness and local theme variation in news articles, and generates a summary for each breakpoint to construct the timeline overview. The overview can provide a clear, global picture for continuously developing news topics, and may benefit news readers in reviewing the topic efficiently. Experiments on real world news topics show the effectiveness of our method.

As future work, we will study the parameter settings in our approach, and optimize the procedures to speed up the algorithm. Furthermore, we plan to collect more data and evaluate our approach on a large variety of news topics.

ACKNOWLEDGMENT

This work is supported by ExxonMobil Research and Engineering Company. We would like to thank the three anonymous reviewers for their valuable comments.

REFERENCES

[1] The Economist, *The News Business: Tossed by A Gale*. 2009.

[2] Wikipedia, *Online Journalism*. 2011.

[3] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron and Yiming Yang, *Topic Detection and Tracking Pilot Study: Final Report*. DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[4] Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat, *Mining Correlated Bursty Topic Patterns from Coordinated Text Streams*. SIGKDD, 2007.

[5] Xiang Wang, Kai Zhang, Xiaoming Jin, and Dou Shen, *Mining Common Topics from Multiple Asynchronous Text Streams*. WSDM, 2009.

[6] Yiming Yang, Tom Pierce, and Jaime Carbonell, *A Study of Retrospective and On-line Event Detection*. SIGIR, 1998.

[7] Yiming Yang, Jaime Q. Carbonell, Ralf D. Brown, Thomas Pierce, Brian T. Archibald, and Xin Liu, *Learning Approaches for Detecting and Tracking News Events*. Intelligent Systems and their Applications, vol 14, pp 32-43, 1999.

[8] Jon Kleinberg, *Bursty and Hierarchical Structure in Streams*. SIGKDD, 2002.

[9] Qiaozhu Mei and ChengXiang Zhai, *Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining*. SIGKDD, 2005.

[10] Alexander Kotov, ChengXiang Zhai, and Richard Sproat, *Mining Named Entities with Temporally Correlated Bursts from Multilingual Web News Streams*. WSDM, 2011.

[11] James Allan, Rahul Gupta, and Vikas Khandelwal, *Temporal Summaries of News Topics*. SIGIR, 2001.

[12] Xiaojun Wan, *TimedTextRank: Adding the Temporal Dimension to Multi-Document Summarization*. SIGIR, 2007.

[13] Fu-ren Lin and Chia-Hao Liang, *Storyline-based Summarization for News Topic Retrospection*. Decision Support Systems, vol 45, pp 473-490, 2008.

[14] Rui Yan, Xiaojun Wan, J. Otterbacher, Liang Kong, Xiaoming Li and Yan Zhang, *Evolutionary Timeline Summarization: A Balanced Optimization Framework via Iterative Substitution*. SIGIR, 2011.

[15] Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li and Yan Zhang, *Timeline Generation through Evolutionary Trans-Temporal Summarization*. EMNLP, 2011.

[16] Griffiths, T. L., Steyvers, M., Blei, D. and Tenenbaum, J., *Integrating Topics and Syntax*. In Advances in Neural Information Processing Systems, vol 17, pp 537-544, 2005.

[17] A. Gruber, Y. Weiss, and M. Rosen-Zvi, *Hidden Topic Markov Models*. In Proceedings of the Conference on Artificial Intelligence and Statistics, 2007.

[18] C. G. Akcora, M. A. Bayir, M. Demirbas, and H. Ferhatosmanoglu, *Identifying Breakpoints in Public Opinion*. 1st KDD Workshop on Social Media Analytics, 2010.

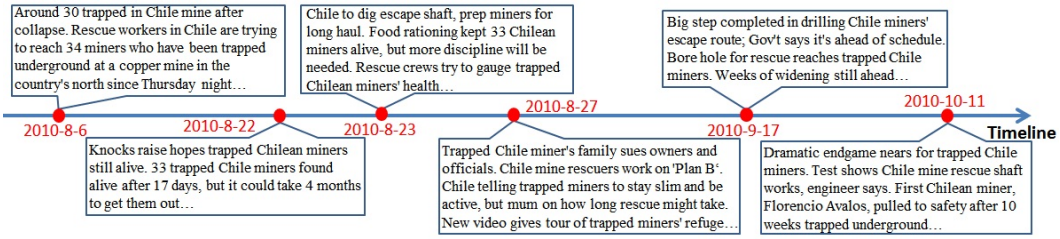

Figure 3. An excerpt from our timeline summary of "Chile Miner".

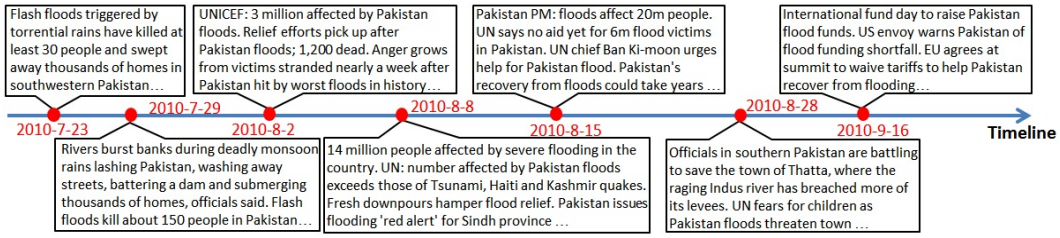

Figure 4. An excerpt from our timeline summary of "Pakistan Floods".

Figure 5. An excerpt from our timeline summary of "Russian Wildfires".

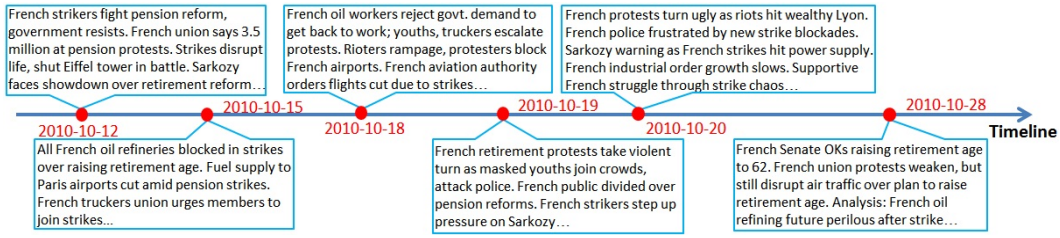

Figure 6. An excerpt from our timeline summary of "French Strikes".

[19] L. R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. IEEE Transactions on Speech and Audio Processing, vol 77, pp 257-286, 1989.

[20] C. Chin Chen, M. Chang Chen, and M. Syan Chen, *An Adaptive Threshold Framework for Event Detection using HMM-based Life Profiles*. ACM Transactions on Information Systems, vol 27, pp 1-35, 2009.

[21] A. P. Dempster and N. M. Laird and D. B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society, vol 39, pp 1-38, 1977.

[22] Dominik M. Endres and Johannes E. Schindelin, *A New Metric for Probability Distributions*. IEEE Transactions on Information Theory, vol 49, pp 1858-1860, 2003.

[23] F. Wilcoxon, *Individual Comparisons by Ranking Methods*. Biometrics, vol 1, pp 80-83, 1945.

[24] Chin-Yew Lin, *ROUGE: A Package for Automatic Evaluation of Summaries*. Workshop on Text Summarization Branches Out, pp 74-81, 2004.