Hu P, Huang ML, Zhu XY. Exploring the interactions of storylines from informative news events. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(3): 502–518 May 2014. DOI 10.1007/s11390-014-1445-6

Exploring the Interactions of Storylines from Informative News Events

Po Hu (胡 珀), Min-Lie Huang* (黄民烈), and Xiao-Yan Zhu (朱小燕), *Member*, *CCF*

State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology Tsinghua University, Beijing 100084, China

E-mail: hup09@mails.tsinghua.edu.cn; {aihuang, zxy-dcs}@tsinghua.edu.cn

Received September 1, 2013; revised March 5, 2014.

Abstract Today's news readers can be easily overwhelmed by the numerous news articles online. To cope with information overload, online news media publishes timelines for continuously developing news topics. However, the timeline summary does not show the relationship of storylines, and is not intuitive for readers to comprehend the development of a complex news topic. In this paper, we study a novel problem of exploring the interactions of storylines in a news topic. An interaction of two storylines is signified by informative news events that play a key role in both storylines. Storyline interactions can indicate key phases of a news topic, and reveal the latent connections among various aspects of the story. We address the coherence between news articles which is not considered in traditional similarity-based methods, and discover salient storyline interactions to form a clear, global picture of the news topic. User preference can be naturally integrated into our method to generate query-specific results. Comprehensive experiments on ten news topics show the effectiveness of our method over alternative approaches.

Keywords text mining, storyline interaction, informative event, coherence, user preference

1 Introduction

The booming of online news industry has brought abundant news resources to the Web. However, the generosity of online news media not only makes news reading more convenient, but also brings a heavy burden to news readers. A reader often finds him/her flooded with tens of thousands of news articles, when inquiring a popular news topic using search engines. As a result, online journalism has become a major source of information overload[1].

To help readers track the events in a continuously developing news topic, online encyclopedias (e.g., Wikipedia) and authoritative news agencies publish manually-edited timelines. A *timeline* is a list of dates in chronological order, and each date has a brief summary of the events. Although a timeline can reduce the burden of readers significantly by listing the "date-event" pairs, it still has several limitations as an ideal news summary. To illustrate the problems, we investigate the timelines compiled by Wikipedia and authoritative news agencies for four popular news topics, i.e., "BP Oil Spill"①-②, "European Debt Crisis"③-④, "Egyptian Revolution"⑤-⑥ and "Libyan Civil War"⑦-⑧. Each news topic has two timelines from Wikipedia and a news agency (i.e., BBC, Wall Street Journal (WSJ), Reuters or CNN) respectively. Table 1 shows the timeline statistics.

In Table 1, the timeline of a news topic has 59 dates on average, and each date has a 50-word summary. Such lengthy content is hard for readers to digest in a timely manner. Considering the reading rate for comp-

Regular Paper

Supported by the National Basic Research 973 Program of China under Grant No. 2012CB316301, the National Natural Science Foundation of China under Grant No. 60803075, the Tsinghua University Initiative Scientific Research Program under Grant No. 20121088071, and the Beijing Higher Education Young Elite Teacher Project.

*Corresponding Author

①http://en.wikipedia.org/wiki/Timeline_of_the_Deepwater_Horizon_oil_spill#2010, Mar. 2014.

②http://www.bbc.co.uk/news/world-us-canada-10656239, Mar. 2014.

③http://en.wikipedia.org/wiki/2000s_European_sovereign_debt_crisis_timeline#2010, Mar. 2014.

④http://online.wsj.com/public/resources/documents/info-EZdebt0210.html, Mar. 2014.

⑤http://en.wikipedia.org/wiki/Timeline_of_the_2011_Egyptian_revolution, Mar. 2014.

⑥http://www.reuters.com/article/2012/05/13/egypt-election-events-idAFL5E8GAECI20120513, Mar. 2014.

⑦http://en.wikipedia.org/wiki/Timeline_of_the_Libyan_civil_war, Mar. 2014.

⑧http://www.cnn.hk/2011/WORLD/africa/08/18/libya.timeline/index.html, Mar. 2014.

Table 1. Timeline Statistics of Four Continuously Developing News Topics

News Topic	Duration	#Dates (News Agency)	Avg. Number of Words/Date (News Agency)	Time Cost (min) (News Agency)
BP Oil Spill	2010.4∼2010.10	76 (Wiki)	42 (Wiki)	16 (Wiki)
		48 (BBC)	56 (BBC)	14 (BBC)
Euro Debt Crisis	2010.1∼2010.11	43 (Wiki)	31 (Wiki)	7 (Wiki)
		62 (WSJ)	43 (WSJ)	14 (WSJ)
Egypt Revolution	2011.2∼2011.12	71 (Wiki)	45 (Wiki)	16 (Wiki)
		36 (Reuters)	29 (Reuters)	6 (Reuters)
Libya Civil War	2011.2∼2011.8	93 (Wiki)	74 (Wiki)	35(Wiki)
		43 (CNN)	49 (CNN)	11 (CNN)
Average		59	49	15

Note: #Dates: the number of time points listed in a topic-specific timeline edited by Wikipedia (Wiki) and the four news agencies respectively; Avg. number of words/date: the average number of words in the summary of a time point listed in a topic-specific timeline edited by Wikipedia (Wiki) and the four news agencies respectively; time cost (min): the estimated time (in minutes) for a reader to comprehend the content of a topic-specific timeline edited by Wikipedia (Wiki) and the four news agencies respectively.

rehension of 200 words per minute[2], we estimate the time needed to comprehend each timeline as listed in Table 1. As a result, a reader is expected to spend 15 minutes without a break to understand the content of a timeline, which is inefficient for the reader to grasp the landscape of the news topic.

The second limitation of a timeline is the lack of storyline relationship. For a complex news topic with intertwined storylines, the non-linear structure of the story development is squashed flat to produce its timeline summary. Fig.1 shows two excerpts from the timeline of "BP Oil Spill" (Fig.1(a)) and the timeline of "European Debt Crisis" (Fig.1(b)). Each event is colored according to the storyline it belongs to.

For both topics in Fig.1, the events in one storyline are separated from each other by events in other storylines. This phenomenon indicates that neighboring events in a timeline may not be coherent. In Fig.1(a), the suspension of well sealing (Jun. 25th) was not due to BP's compensation (Jun. 16th) or the drilling moratorium (Jun. 22nd). In Fig.1(b), the EU's bailout (May 10th) was not caused by the violent protests ahead of it. However, storylines do have interactions with each other through certain events. In the BP case, the static kill (Jul. 19th) in *Well sealing* led to the Gulf reopen (Jul. 22nd) in *Environment*. In the Euro Debt case, the austerity measures (Apr 29th.) in *Greek government* triggered the violent protests in *Greek riot* and the bailout in *EU rescue*. Since a timeline organizes the events only by their time stamps, it is hard for readers to identify the relationship of events (and storylines) from the distractive and incoherent event threads, let alone interpret the development of the whole story.

Finally, a timeline is a static summary that cannot be adjusted for user preference. If a reader is interested in certain aspect of a news topic (e.g., BP's compensa-

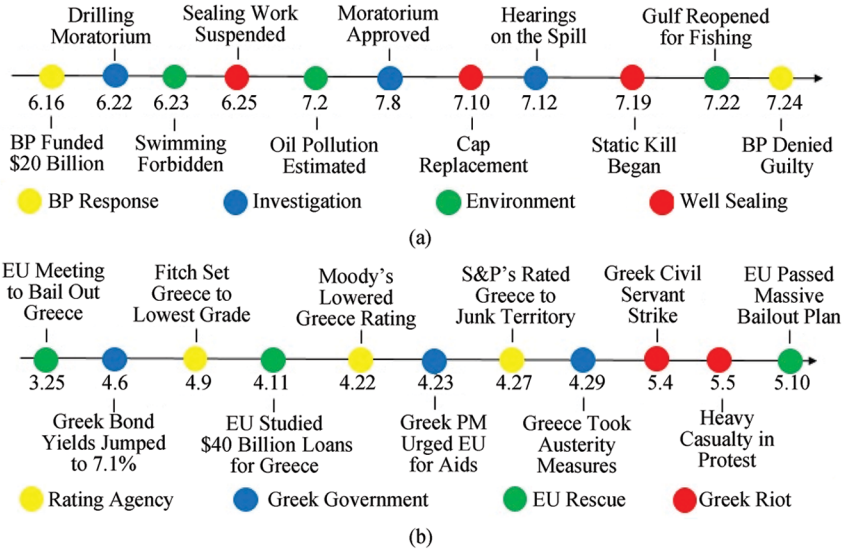


Fig.1. Two excerpts from (a) the timeline of "2010 BP Oil Spill" by Wikipedia and (b) the timeline of "2010 European Debt Crisis" by Wall Street Journal.

tion for the spill), the only way for him/her is to do a keyword search in the timeline summary. However, the inherent ambiguity of words can make the results unsatisfactory. For example, although compensation is a key issue in "BP Oil Spill", the term *compensation* is not found in the timeline by Wikipedia. Instead, alternative words are used to summarize the corresponding events. Furthermore, some events of the desired aspect are not included in a timeline, as editors need to trade off among various aspects to keep the summary unbiased.

In this paper, we address the above limitations in news timeline and study a novel text mining problem, i.e., exploring the interactions of storylines in a news topic. Storyline interactions can indicate key phases of a news topic, and reveal the latent connections among various aspects of the story. Two storylines interact if there exists two events, each belonging to one of the storylines, have a strong correlation (e.g., cause and effect) with each other. The events bridging the storylines are termed *informative events*, since they are crucial for outlining the structure of the evolving stories. Unlike traditional similarity-based methods, our approach addresses the coherence between news articles, and can effectively discover informative events from the coherence graph. Based on the informative events, salient storyline interactions are extracted to form a structured overview of the news topic. User preference can be naturally integrated into our method to generate query-specific results, thus help readers navigate the story development at a global as well as focused view.

The rest of the paper is organized as follows. We survey related work in Section 2. Section 3 formulates the problem and presents our methodology. Details of the approach are described in Section 4. Section 5 analyzes the experimental results. Section 6 concludes the paper.

2 Related Work

To the best of our knowledge, storyline interaction analysis is a novel topic in the area of topic detection and tracking (TDT)[3]. Different from automatic news timeline generation, which has been extensively studied in recent years[4-7], our work aims to reveal the latent connections among storylines with their implications, and explicitly show the relationship of information nuggets in a timeline.

Our work also differs from the studies of TimeML (e.g., [8-9]). TimeML is a markup language that identifies time and events by annotating the temporal expressions in text. An event in TimeML is more of a specific action (e.g., perception or state) than a news event that may comprise of multiple actions. In addition, the temporal precedence of events in TimeML is inferred by the Reichenbach tense analysis[10] and Allen's temporal logic[11], both are linear logics that represent events in a linear manner, thus are not suitable to analyze the non-linear storyline interactions.

Furthermore, our research is different from temporal topic evolution analysis[12-17], which analyzes the evolution of topics with their popularities over time. Cui *et al.*[14] and Gao *et al.*[15] worked on revealing the connections among topics discovered from the text data, and they focused on how one topic splits into multiple topics, and how multiple topics merge into one topic. An incremental HDP (hierarchical Dirichlet process)-based approach[18] was proposed that extracts a set of topics from a text collection and models the splitting/merging patterns among evolving topics. The method also identifies events that triggered the splitting/merging patterns. Finally, the topic evolution process is visualized, including topic strength, content, and splitting/merging relationship. Unlike the above work that detects the topic transitions from one to another (e.g., topic splitting and merging), our research focuses on the interactions of multiple storylines which are developed simultaneously in the time span of a news topic. In addition to topic splitting and merging, storyline interaction is also crucial to characterize the connections among topics, yet it has been much less explored in the research field.

Our research is relevant to some previous work that studies the relationship of news articles. Nallapati *et al.*[19] identified events with their dependencies in a news topic. An event is defined as an exclusive news cluster based on word similarity, and each article can only belong to one event. An event is dependent on another event if the average similarity of their articles is above a given threshold. Mei *et al.*[20] discovered evolutionary theme patterns in a news topic. The news stream is sliced into time intervals, and themes are extracted from each interval using a probabilistic mixture model. Two themes in different intervals are connected to form an evolutionary pattern if their word distributions are highly similar. Choudhary *et al.*[21] studied actor transformations in a news topic. An actor is defined as a word or a phrase that occurs repeatedly in the news stream. News articles with high textual similarities are connected to each other, and the strength of a connection is determined by the co-occurrence of actors in the articles (i.e., actor transformations). The articles with salient actor transformations are selected to form a summary of the news topic.

Unlike the research based on word similarity, our method utilizes coherence to determine the correlations of news articles in different storylines. The proposed

method is more effective than the above work in the experiments.

Finally, Shahaf *et al.*[22] generated a metro map to visualize the progress of a news topic. Although we share the same motivation with theirs (i.e., to help readers understand the story development), the research focus is different. The metro map illustrates the topic progress by organizing both key events and side stories into several event threads, while our method summarizes the topic progress through storyline interactions. In addition, the algorithm used in [22] is very complicated (and the source code is not open to public), and the performance is evaluated through a user survey. In contrast, our framework is much easier to implement, and the performance is quantitatively analyzed by gold standard derived from Wikipedia.

3 Problem Formulation

Suppose a reader wants to review a news topic Q, which has a stream of relevant news articles $\mathcal{D} = \{D_t|t = 1, \ldots, T\}$, where $D_t \subset \mathcal{D}$ is a news collection of Q published at date t.

We follow the definition in TDT that an *event* is a particular thing that happens at a specific time and place[3]. A news article $d \in \mathcal{D}$ narrates an event in Q. In journalism, the Five Ws (i.e., When, Where, Who, What and Why) constitute the key elements of d, and cover the time, place, actors and their actions in the event[23]. The events in a news topic are focused on different aspects of the story, and we define a storyline as:

Definition 1 (Storyline). *A storyline S in a news topic Q is a chain of events that characterize a certain aspect of Q and involve the same set of actors and places.*

The above definition of storyline is consistent with its original meaning in dramatic discourse studies, where a storyline (a.k.a. narrative thread) refers to the writing to center the part of the story in the action or experience of specific sets of characters, thus the narrative threads experienced by different characters are woven together to form the plot of a play[24]. Fig.1 shows excerpts of four storylines in two news topics respectively. For example, the storyline *Well sealing* in Fig.1(a) is focused on the efforts to seal the leaking oil well, and the events in the storyline involve the same actors (e.g., BP, the U.S. Coast Guard) and places (e.g., the Gulf of Mexico, the Macondo Prospect Oil Field).

A news topic may have multiple storylines depending on the complexity of its nature. The storylines are not isolated from each other; one storyline can influence or be influenced by other storylines through certain events. An interaction of storylines is defined as:

Definition 2 (Interaction). *An interaction of two storylines S_i and S_j in a news topic Q occurs, if an event $\mathcal{E}_i$ in S_i has a strong correlation (e.g., cause and effect) with an event $\mathcal{E}_j$ in S_j.*

We assume that each event is focused on a single aspect of the story, thus it can only belong to one storyline in Q. In the above definition of storyline interaction, we term $\mathcal{E}_i$ and $\mathcal{E}_j$ *informative news events*, since such events are the bridges of different storylines, and represent the informative parts of Q that best characterize the structure of the topic progress.

Storyline interactions are crucial to a news topic, as informative events can alter the track of the interacted storylines, and influence the development of the whole story. Fig.2(a) shows the interactions of four storylines in "2012 Senkaku (Diaoyu) Islands Dispute". After calming down from the first wave of anti-Japanese protests (since Aug. 19th), the storyline *Chinese protests* broke out again (Sept. 13th) triggered by the informative event Islands nationalization (Sept. 10th) in *Japanese government*.

The key problem in storyline interaction analysis is measuring the correlation between two events. In previous studies (e.g., [19, 21]), an event is a cluster of similar documents, or a selected document in the cluster. The correlation between two events is determined by word similarity (e.g., cosine similarity) of the documents. Such similarity-based measurements can cause serious word mismatch problem, i.e., only those articles with the same wordings can be connected, while the relationship of different events is missed.

In this paper, the correlation between two events is determined by the *coherence* of news articles. Compared with word similarity, coherence is a high level concept of content consistency. Similar news articles are coherent, while coherent articles may not be similar in words. In Fig.2(a), the follow-up reports in *Chinese protests* share much words with the news of factory shutdown (Sept. 15th), and those articles are indeed coherent since they are in the same storyline. In contrast, the news of protest restriction (Sept. 16th) in *Chinese government* is much less similar to the news of factory shutdown, yet they are still coherent since the latter is a major cause of the former event.

Each event in Q is represented by features of the Five Ws extracted from the document text. For a news article $d \in D_t$ that narrates an event $\mathcal{E}$, the publication date t is the time when $\mathcal{E}$ happens. The Location-class entities in d constitute the place where $\mathcal{E}$ happens. The Person-class and Organization-class entities in d are the actors in $\mathcal{E}$. The subtopic distribution of d characterizes the focused aspect of $\mathcal{E}$, and is explained in detail in Subsection 4.1.1.

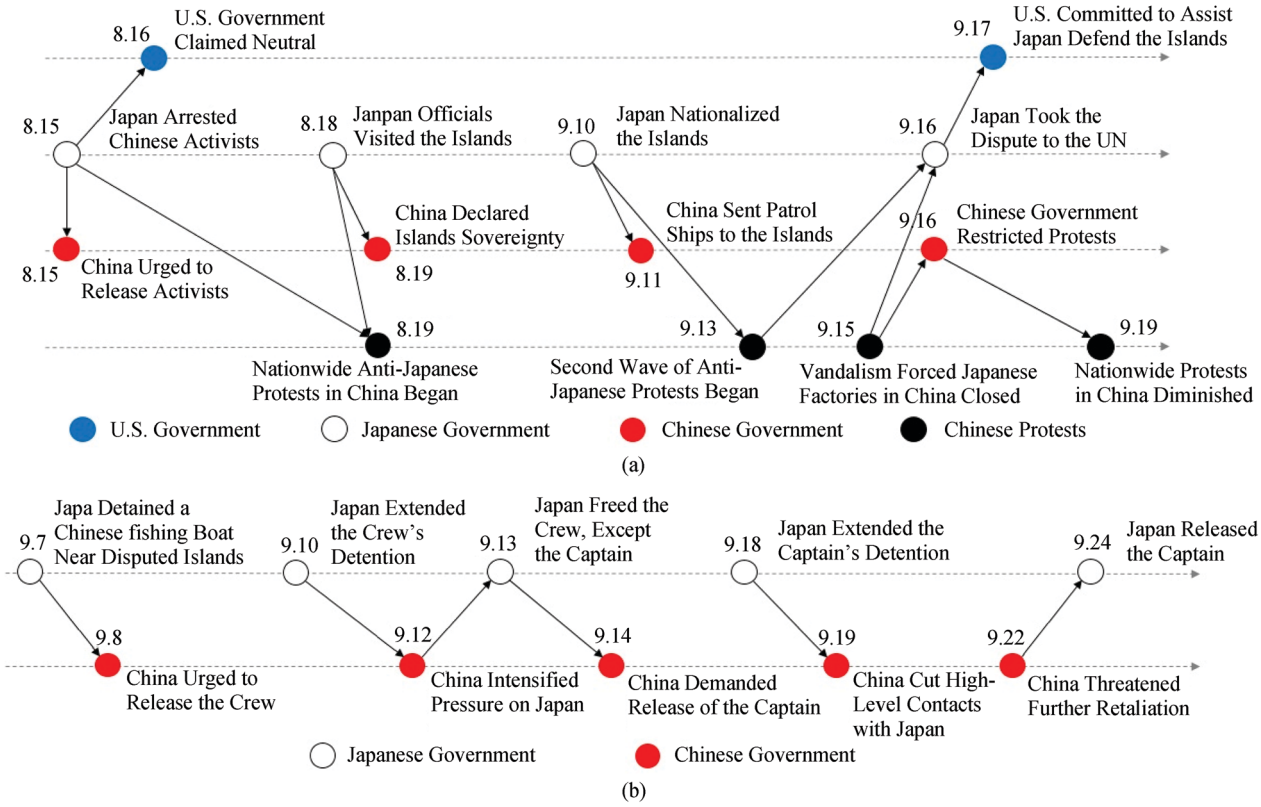


Fig.2. Two excerpts of the structured overview of (a) "IslandsClash 2012⋆" and (b) "Islands Clash 2010†". The title of each storyline is labeled by the main actors or actions in the storyline, and arrows indicate the temporal order of news. Events not involved in the storyline interactions are not shown in the graph.

The storylines in Q are also differentiated by the entity and subtopic features. The entity features (i.e., the three classes of named entities) annotate the *dramatis personae* in the storyline, while the subtopic distributions specify the focused aspect of the storyline.

Based on the above features, the coherence between a pair of news articles d_i and d_j is determined by three factors, i.e., time continuity, entity relatedness, and subtopic consistency. Time continuity measures the time distance between the events in d_i and d_j. Entity relatedness calculates the affinity of the entity features in the events. Subtopic consistency matches the subtopic distributions in the events. We have designed effective measurements for the three coherence factors, which form the foundation of storyline interaction analysis.

4 Methodology

Fig.3 demonstrates the framework of the proposed method, starting from the data collection step to the final output step. There are three main steps in the framework:

Step 1: extract the entity features and the subtopic features from the document text, and build the coherence graph;

Step 2: identify the informative events from the coherence graph through random walk, with user preference integrated;

Step 3: discover salient storyline interactions based on the informative events, and generate a structured overview of the news topic.

4.1 Coherence Graph Construction

Three classes of named entities (i.e., Person, Organization and Location) are extracted from the news corpus $\mathcal{D}$ using the Stanford NER tools[25]. Each entity is then hyphenated like a unigram word (e.g., "United Nations" to *United-Nations*). Finally, all words are lowercased with stop words removed in each document in $\mathcal{D}$.

4.1.1 Subtopic Consistency Factor

A generative probabilistic mixture model[26] is used to discover the latent subtopics. Suppose there are K subtopics $\mathcal{Z} = \{z_k|k = 1, \ldots, K\}$ and a background topic z_B in $\mathcal{D}$. A subtopic z_k is a probabilistic distribution of words in the vocabulary $\mathcal{W}$ of $\mathcal{D}$; that is, z_k governs the multinomial distribution of words $\{p(w|z_k)|w \in \mathcal{W}\}$ s.t. $\sum_{w \in \mathcal{W}} p(w|z_k) = 1$. A document d

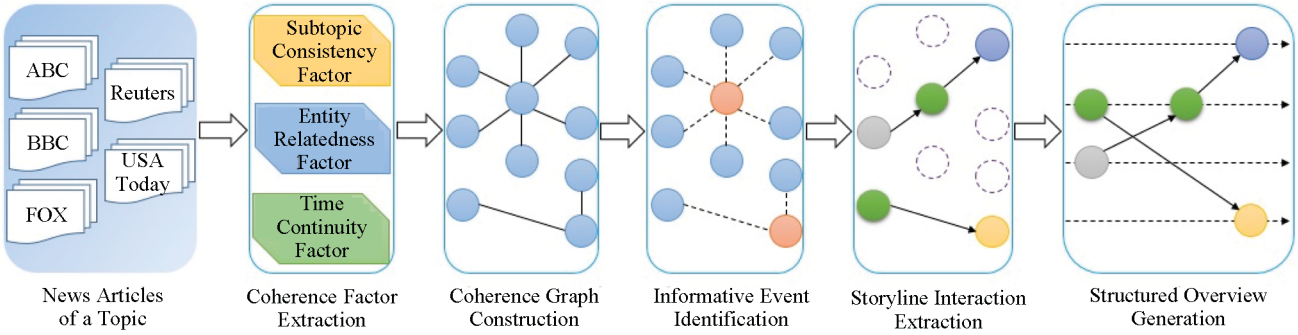


Fig.3. Framework of the proposed method.

is a probabilistic distribution of subtopics $\{p(z_k|d)|z_k \in \mathcal{Z}\}$ s.t. $\sum_{z_k \in \mathcal{Z}} p(z_k|d) = 1$. A word w in a document d is sampled according to the following probability:

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

where λ_B is the weight for z_B, and is selected ad hoc. The occurrence of a word w given z_B is estimated as $p(w|z_B) = \frac{\sum_{d \in \mathcal{D}} c(w,d)}{\sum_{w' \in \mathcal{W}} \sum_{d \in \mathcal{D}} c(w',d)}$, where $c(w,d)$ is the number of occurrences of w in d. The background topic z_B is formed with high-frequent but low-informative words (i.e., domain stop words). The K subtopics are more discriminative and meaningful due to the introduction of z_B. The model's parameters, $\{p(z_k|d)\}$ and $\{p(w|z_k)\}$, can be estimated by using the EM algorithm[27].

For a pair of news articles d_i and d_j in $\mathcal{D}$, the distance of their subtopic distributions is measured by the square root of the Jensen-Shannon divergence[28], which is a metric defined as:

$$SD(d_i, d_j) = \sqrt{\begin{array}{l} \frac{1}{2} \sum_{k=1}^{K} p(z_k|d_i) \log \frac{p(z_k|d_i)}{m_{z_k}} + \\ \frac{1}{2} \sum_{k=1}^{K} p(z_k|d_j) \log \frac{p(z_k|d_j)}{m_{z_k}}, \end{array}} \quad (2)$$

where

$$m_{z_k} = \frac{1}{2}[p(z_k|d_i) + p(z_k|d_j)].$$

The subtopic consistency factor of d_i and d_j is defined as:

$$SC(d_i, d_j) = 1 - SD(d_i, d_j). \quad (3)$$

In log base 2, $SD(d_i, d_j) \in [0, 1]$, thus $SC(d_i, d_j)$ is in the range of [0, 1]. More similar subtopic distributions in d_i and d_j will generate higher values of $SC(d_i, d_j)$, which indicate that d_i and d_j are more likely to be focused on the same aspects of the story.

4.1.2 Entity Relatedness Factor

The entity relatedness factor of d_i and d_j is determined by the affinity of the entities in the document pair. Suppose NE_i is the set of the three classes of named entities in d_i. A naive method to determine the entity relatedness factor is based on the entity overlap:

$$Jaccard(d_i, d_j) = \frac{|NE_i \bigcap NE_j|}{|NE_i \bigcup NE_j|}, \quad (4)$$

where $Jaccard(d_i, d_j) \in [0, 1]$ is the ratio of the entities shared by d_i and d_j. This simple method considers two entities are related only if they are the same word, and totally ignores the semantic relatedness of different entities, thus it cannot meet our needs.

Instead, we use the normalized pointwise mutual information (NPMI) to measure the affinity of an entity pair $e_i \in NE_i$ and $e_j \in NE_j$, which is defined as:

$$NPMI(e_i, e_j) = \frac{PMI(e_i, e_j)}{-\log p(e_i, e_j)}, \quad (5)$$

where

$$PMI(e_i, e_j) = \log \frac{p(e_i, e_j)}{p(e_i)p(e_j)}.$$

In log base 2, $NPMI(e_i, e_j) \in [-1, 1]$. Higher values of $NPMI(e_i, e_j)$ indicate the two entities are more related to each other, and $NPMI(e_i, e_j)$ is 0 if e_i and e_j are completely independent.

A classical method utilizes entity statistics to interpret $p(e_i)$ and $p(e_i, e_j)$, in which $p(e_i)$ is the term frequency of e_i in $\mathcal{D}$, and $p(e_i, e_j)$ is the co-occurrence frequency of e_i and e_j in $\mathcal{D}$. Although this method can link different entities that frequently co-occur in the corpus, it still suffers from the word mismatch problem, especially when the documents are collected from different sources with their own writing styles.

In this paper, we interpret the entity relatedness on the topic level, and determine $p(e_i)$ and $p(e_i, e_j)$ as:

$$p(e_i) = \sum_{k=1}^{K} p(e_i|z_k)p(z_k), \quad (6)$$

$$p(e_i, e_j) = \sum_{k=1}^{K} p(e_i, e_j|z_k)p(z_k), \quad (7)$$

$$p(z_k) = \sum_{i=1}^{|\mathcal{D}|} p(z_k|d_i)p(d_i), \quad (8)$$

where $p(d_i) = \frac{1}{|\mathcal{D}|}$. In the probabilistic mixture model discussed in Subsection 4.1.1, two words e_i and e_j are conditionally independent given the subtopic z_k, thus $p(e_i, e_j|z_k) = p(e_i|z_k)p(e_j|z_k)$. The proposed method measures the entity relatedness through the parameters in the mixture model, and can link semantically related entities without entity disambiguation.

Finally, the entity relatedness factor of d_i and d_j is defined as:

$$ER(d_i, d_j) = \frac{\sum_{e_i \in NE_i} \sum_{e_j \in NE_j} NPMI(e_i, e_j)}{|NE_i| \times |NE_j|}, \quad (9)$$

where $ER(d_i, d_j) \in [-1, 1]$. Higher values of $ER(d_i, d_j)$ indicate that d_i and d_j are more coherent on the entity features.

4.1.3 Time Continuity Factor

The time continuity factor of d_i and d_j is determined by the time distance between the two documents, and is measured by the Gaussian window function as:

$$F(\Delta t) = e^{\frac{-\Delta t^2}{2\sigma^2}}, \quad (10)$$

where $\Delta t = t_i - t_j$, σ is the decay rate of the Gaussian window, and $F(\Delta t) \in [0, 1]$. Higher values of $F(\Delta t)$ indicate that d_i and d_j are more coherent on the time dimension.

4.1.4 Coherence Graph Construction

The coherence score of d_i and d_j is a combination of the three coherence factors as:

$$C_{i,j} = F(\Delta t)(\mu SC(d_i, d_j) + (1 - \mu)ER(d_i, d_j)), \quad (11)$$

where $\mu \in [0, 1]$ is used to balance the influence of the entity features and the subtopic features. For any coherent document pair d_i and d_j, the coherence factors (i.e., $F(\Delta t)$, $SC(d_i, d_j)$ and $ER(d_i, d_j)$) should all exceed zero. Thus, $C_{i,j} \in (0, 1]$ for all coherent document pairs.

Based on the notion of coherence formulated above, we build the coherence graph which is defined as:

Definition 3 (Coherence Graph). *A coherence graph $G = (V, E)$ is an undirected graph built on the corpus $\mathcal{D}$ of a news topic Q, where each vertex $v \in V$ represents a news article $d \in \mathcal{D}$. Two vertices v_i and v_j are connected by an undirected edge $e_{i,j} \in E$, if the coherence score $C_{i,j}$ is higher than a given threshold ϵ.*

In the above definition, we treat coherence as a constraint subjected to the threshold of acceptance, which is consistent with the observations from psychologists that coherence is a constraint satisfaction in human cognitive process[29]. The structure of the coherence graph G is determined by the coherence scores, which are determined by the three coherence factors (i.e., time continuity, entity relatedness, and subtopic consistency) that characterize the key elements (i.e., the Five Ws) of news articles[23].

When calculating the coherence score as in (11), a linear combination is firstly made from the entity relatedness factor and the subtopic consistency factor, which constitute the textural features (i.e., Where, Who, What, and Why) of news, then the result is combined with the temporal feature (i.e., When) of news. Thus, (11) provides an intuitive way to combine the three factors while reserving the distinction between the textural features and the temporal feature of news articles.

Another advantage of (11) is that it permits fast construction of the coherence graph. The complexity of building the coherence graph G is $\mathcal{O}(|\mathcal{D}|^3|\mathcal{W}|^2K)$, which is infeasible for a complex topic with tens of thousands of news articles. To accelerate the building process, the coherence factors of two documents are calculated only if their time distance is not too far. We have observed that once $|\Delta t|$ exceeds 1.8σ in the Gaussian window, the time continuity factor $F(\Delta t)$ drops below 0.2, and the document pair can hardly be coherent in our corpus. Thus, the coherence graph G is built within a sliding window in which the time distance of any document pair $|\Delta t| \leqslant 1.8\sigma$.

4.2 Informative Event Identification

Informative events connect different storylines, and constitute the structured overview of the news topic. In this paper, we aim to not only discover informative events in the development of the whole story, but also identify informative events that are biased toward user preference (which is represented by query words). In the former case, the desired events are *informative* in the scope of the entire news topic. In the latter case, the desired events are *informative* with regard to the storylines that are interesting for a reader. To better incorporate user preference than traditional keyword search, we propose a novel method that integrates

the subtopic features into the topic-sensitive PageRank algorithm[30], and discover the informative events from the coherence graph G in either a global or a focused view. Fig.4 is an illustration of the proposed method. Orange nodes denote documents of the informative events, and blue nodes denote non-informative documents. If no user preference is given, the algorithm finds the informative events in the global scope (i.e., document A and C in Fig.4(a)). If a user query q is issued, the algorithm then finds the informative events that are biased towards q. Therefore, document B which is in the user-interested storyline as A is selected in Fig.4(b), while document C which is in other storyline is not.

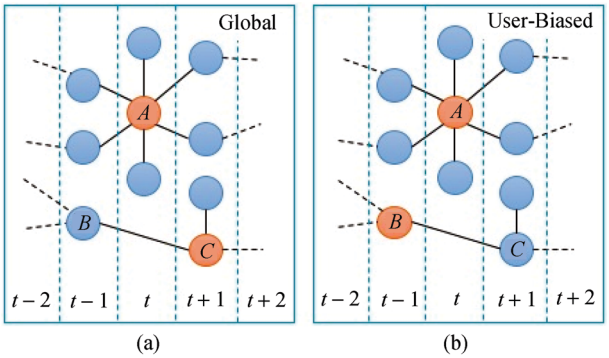


Fig.4. Illustration of the proposed method.

Different from previous studies (e.g., [19-21]) that organize the documents into directed acyclic graphs (DAGs), the coherence graph G is an undirected graph, and the random walker can move bidirectionally between a pair of linked vertices in G. The DAG is not suitable to perform PageRank-style algorithms, since the highest PageRank scores will be assigned to the vertices at the end of each storyline (e.g., $D_T \subset \mathcal{D}$), and miss the real informative events. Besides, the direction of an edge is uncertain for some articles that are published on the same date.

Each vertex (corresponding to a document d) has K PageRank scores $\{PR(z_k, d)|k = 1, \ldots, K\}$, and score $PR(z_k, d)$ reflects the importance of d to subtopic z_k. The PageRank score $PR(z_k, d)$ is initialized as $\frac{1}{K|\mathcal{D}|}$, and is calculated in an iterative process:

$$PR_{i+1}(z_k, d) = (1-\alpha)\boldsymbol{P}_{z_k,d} + \alpha \sum_{d' \in B_d} \frac{PR_i(z_k, d')}{L(d')}, \tag{12}$$

where $\alpha = 0.85$ is the damping factor. B_d is the set of vertices linking to the vertex of d, and $L(d')$ is the degree of the vertex of d'. $\boldsymbol{P}_{z_k}$ is the damping vector for z_k, and is initialized as an uniform column vector, i.e., $\boldsymbol{P}_{z_k} = \left[\frac{1}{|\mathcal{D}|}\right]_{|\mathcal{D}|\times 1}$.

Equation (12) treats the edges of d equally in d's PageRank score, and ignores the corresponding coherence scores once they exceed the threshold ϵ. In other words, coherence is a constraint, instead of the object in informative event identification. There are two reasons for it. Firstly, according to the research in psychology, coherence is a constraint satisfaction in human cognitive process, and people are satisfied once the degree of coherence reaches the threshold of their acceptance[29]. Secondly, maximizing coherence in PageRank by incorporating the coherence scores in (12) will make the algorithm biased towards the most similar documents, thus will produce highly redundant results.

The final score of the vertex of d is a combination of its PageRank scores:

$$Score(d) = \sum_{k=1}^{K} \lambda_k PR(z_k, d), \tag{13}$$

where λ_k is the weight for subtopic z_k w.r.t. the user preference. If no user query is issued, then λ_k is set as 1. If a reader is interested in certain aspects of the news topic, and issues a query $q = \{w_i|w_i \in \mathcal{W}\}$[⑨], then the weight λ_k for subtopic z_k w.r.t. q is determined as:

$$\lambda_k = p(z_k|q) = \frac{p(z_k)p(q|z_k)}{p(q)}, \tag{14}$$

where $p(q)$ is a constant once q is issued, thus $\lambda_k \propto p(z_k)p(q|z_k)$. $p(q|z_k) = \prod_{w_i \in q} p(w_i|z_k)$, and $p(z_k)$ is obtained by (8).

Accordingly, the damping vector $\boldsymbol{P}_{z_k}$ w.r.t. q is determined as:

$$\boldsymbol{P}_{z_k} = [p(d|z_k)]_{|\mathcal{D}|\times 1} = \left[\frac{p(z_k|d)p(d)}{p(z_k)}\right]_{|\mathcal{D}|\times 1}. \tag{15}$$

By (8), $\boldsymbol{P}_{z_k}$ is simplified as:

$$\boldsymbol{P}_{z_k} = \left[\frac{p(z_k|d)}{\sum_{i=1}^{|\mathcal{D}|} p(z_k|d_i)}\right]_{|\mathcal{D}|\times 1}. \tag{16}$$

When the PageRank algorithm is converged, we rank all vertices in G by their scores calculated in (13). The top-ranked vertices are selected to represent the informative events in the news topic.

4.3 Storyline Interaction Extraction

Each top-ranked vertex in G represents an informative event in the news topic, and we analyze the micro-

⑨All the query words are lowercased, and the words not appeared in the news corpus $\mathcal{D}$ are filtered out.

environment of such vertices to extract the salient storyline interactions. Algorithm 1 is the storyline interaction mining algorithm.

Algorithm 1. Storyline Interaction Mining
Input: coherence graph G; document list sorted by the PageRank score $\mathcal{D}_{\text{sort}}$; maximum number of storyline interactions N;
Output: a set of storyline interactions $\mathcal{I}$;
1: $\mathcal{I} \Leftarrow \varnothing$; $\mathcal{D}_I \Leftarrow \varnothing$;
2: **for** each $d \in \mathcal{D}_{\text{sort}}$ **do**
3: $z_{dom(d)} \Leftarrow \arg\max_{1 \leqslant k \leqslant K}\{p(z_k|d)\}$;
4: **if** $\nexists d_0 \in \mathcal{D}_I$ s.t. $t_d = t_{d_0}$ and $z_{dom(d)} = z_{dom(d_0)}$ **then**
5: **for** each d' that is connected to d in G **do**
6: **if** $z_{dom(d)} \neq z_{dom(d')}$ **then**
7: **if** $\nexists d'_0 \in \mathcal{D}_I$ s.t. $(d, d'_0) \in \mathcal{I}$ and $t_{d'} = t_{d'_0}$ and $z_{dom(d')} = z_{dom(d'_0)}$ **then**
8: $\mathcal{I} \Leftarrow \mathcal{I} + (d, d')$;
9: $\mathcal{D}_I \Leftarrow \mathcal{D}_I + d$;
10: $\mathcal{D}_I \Leftarrow \mathcal{D}_I + d'$;
11: **end if**
12: **if** $|\mathcal{I}| = N$ **then**
13: **return** $\mathcal{I}$;
14: **end if**
15: **end if**
16: **end for**
17: **end if**
18: **end for**

After an important event happens, different news agencies publish their own reports about the event. If a news article narrating an informative event has a high PageRank score, those reports on the same event by other agencies are also likely to receive high PageRank scores, which should be removed from the top-ranked documents to avoid redundancy. A traditional method is to cluster the news articles based on word similarity (e.g., cosine similarity of TF-IDF vectors). However, as pointed out by [19], the parameters of the clustering algorithm are hard to choose for different news topics, which can affect the clustering performance tremendously.

Instead, we assume that a storyline can have at most one informative event in any given date, and the assumption holds in our dataset that covers 10 popular news topics in recent years (see Subsection 5.1). For example, on the day 2012/9/16 in the topic "2012 Senkaku (Diaoyu) Islands Dispute" (see Fig.2), the only informative event in the storyline *Chinese government* is the Protest restriction, and there is no other key event happened in the same storyline on the same day. However, there could exist informative events on the same day in other storylines, such as Japan's proposal to the UN (Sept. 16th) in the storyline *Japanese government*.

For a news article d of an informative event $\mathcal{E}$, the *dominant subtopic* $z_{dom(d)}$ (i.e., the subtopic with the highest probability in $\{p(z_k|d)|z_k \in \mathcal{Z}\}$) distinguishes the storyline $\mathcal{E}$ belongs to. Once a news article d is processed to extract storyline interactions, all the other documents on the same date with the same dominant subtopic as $z_{dom(d)}$ are considered to report on the same informative event, and are removed from the top-ranked documents to avoid redundancy (refer to step 4 in Algorithm 1).

For a top-ranked vertex (corresponding to document d) in G, two classes of vertices are linked to it. The first class is the documents in the same storyline with d, such as the follow-up reports of the event in d. The second class is the documents in which the events are highly correlated (e.g., cause-effect relationship) with the event in d, and those documents represent the interactions of different storylines. Distinction between the two classes of vertices is based on the dominant subtopic in each document. Only those documents connected to d with different dominant subtopics are extracted to represent the storyline interactions (refer to steps 5∼11 in Algorithm 1), and the interactions that have already been extracted previously are ignored (refer to step 7 in Algorithm 1). In this way, the editorials and comments on the Protest restriction (Sept. 16th) in the storyline *Chinese government* are filtered out, while the vandalism and violence in the storyline *Chinese protests* that led to the protest restriction are revealed (see Fig.2).

Algorithm 1 is stopped once the discovered storyline interactions have reached the desired amount (refer to steps 12∼13 in Algorithm 1). The maximum number of storyline interactions N is user-specified, and has a zooming effect on the structured overview. Increasing N will disclose richer interaction details in the topic development, while decreasing N will only exhibit the most significant storyline interactions. Finally, we organize the storyline interactions chronologically to form a structured overview of the news topic. Fig.2 shows the excerpts of the overview of two news topics in our dataset.

5 Experiment

5.1 Dataset

Existing evaluation frameworks such as TDT cannot be used as benchmark in our task. For example, the TDT 2004 dataset[⑩] contains 278 109 news articles of

[⑩] http://www.itl.nist.gov/iad/mig/tests/tdt/2004/workshop.html, Apr. 2014.

250 topics. Each topic is labeled with its actors, places, and a timeline summary. Although the dataset is qualified for the tasks such as story segmentation and first story detection, it is not applicable to storyline interaction analysis. The reason is that the 250 topics are not related to each other at all after manual inspection.

We therefore construct our own dataset. Ten popular news topics are selected covering politics, finance, conflicts, and disasters in recent years. For each topic, we download the news articles from five authoritative news agencies (i.e., ABC, BBC, FOX, Reuters, and USA Today), through keyword search with time constraint in their websites. For each topic, we check the number of dates in the timeline from Wikipedia as an indicator of the topic complexity. Table 2 shows the dataset statistics.

In Table 2, the ten topics are divided into two categories: four *simple topics*, each focusing on a single geographic location for a short time period, marked by †; six *complex topics*, each spanning multiple countries for a longer time period, and the corresponding timeline has much more dates than that of a simple topic, marked by ⋆. Algorithms are tested on both categories to evaluate their performances in various scenarios that news readers may face.

Finally, we investigate the volume of the dataset in real world applications. The proposed approach can be integrated into a news search engine, in which the output of the search engine (i.e., topic-relevant news articles) is used as the input of the storyline interaction analyzing algorithm. We therefore check the number of news articles retrieved by Google news archives for each topic, by querying the keywords in the search engine with time constraints. As shown in Table 2, for most of the news topics, the dataset has the same order of magnitude of the number of documents retrieved by the search engine. Thus, the dataset can be used to simulate real world applications when the proposed algorithm is integrated into a news search engine.

5.2 Gold Standard

Wikipedia is chosen as the source of the ground truth, considering the content accuracy, writing quality, and topic coverage of its articles. For each topic in our dataset, the corresponding articles in Wikipedia are read by domain experts to construct the gold standard of storyline interactions.

In practice, our annotators have found that the informative events in a topic are hard to label, while correlations among the events are relatively easy to identify from Wikipedia articles. For example, in the article of "Islands Clash 2012", some narrations such as "The Ministry of Foreign Affairs has strongly protested *with regard to* the landing of activists on the islands" and "Many Japanese businesses in China were shuttered *in reaction to* the protests", have clearly shown the strong correlation between two events. Therefore, we ask annotators to label the correlated event pairs from Wikipedia articles as the first step to construct the gold standard.

The second step is to check whether a pair of correlated events come from different storylines to make sure they are qualified to represent storyline interactions. The heading or subheading of each chapter in Wikipedia articles is used to label a storyline in the topic after manual inspection[11]. For example, the chapters in the article of "BP Oil Spill" include *Well sealing*, *Environment*, *Investigation*, *Litigations*, *Economic impact*, etc.; each labels a unique storyline in the topic. The main article of each chapter is used to label the events in the corresponding storyline. For each topic, only those event pairs that come from different storylines are added to the gold standard.

Table 2. Dataset Statistics

Topic	Duration	Brief Description	$\#\mathcal{D}$	$\#\mathcal{P}$	$\mathcal{O}$
BP Oil Spill⋆	2010.4∼2010.10	The largest oil spill accident in the history	2 272	76	10^3
Euro Debt Crisis⋆	2010.1∼2010.11	The worst financial crisis in the European Union	2 418	43	10^3
Haiti Earthquake⋆	2010.1∼2010.3	The second deadliest earthquake in the history	2 072	65	10^3
Arab Spring⋆	2010.12∼2011.4	Revolutions spread across Arab countries	2 279	40	10^3
Iceland Ash⋆	2010.3∼2010.6	European air traffic suspended by volcanic ash	1 384	44	10^3
Islands Clash 2012⋆	2012.4∼2012.11	Japan nationalized the disputed islands	1 209	28	10^3
Thailand Riot†	2010.3∼2010.5	Violent political protests held by the Red-Shirts	642	14	10^3
Chilean Miner†	2010.8∼2010.10	The longest mine accident rescue in the history	523	15	10^2
Islands Clash 2010†	2010.9∼2010.11	Boat collision near the disputed islands	242	16	10^2
Russian Fire†	2010.7∼2010.8	Extensive wildfires caused by extreme heat waves	216	13	10^2

Note: $\#\mathcal{D}$ is the number of documents in a news topic; $\#\mathcal{P}$ is the number of dates in the timeline by Wikipedia; $\mathcal{O}$ is the order of magnitude of the number of documents retrieved by Google news search engine.

[11] Some chapter headings (e.g., the Background) are not used to label storylines, and subheadings with similar content (e.g., Volunteer efforts and Relief efforts) are merged into one storyline in the gold standard.

Three domain experts are solicited to build the gold standard. Through the investigation of event correlation and storyline membership, a valid event pair needs at least two experts' agreements. Finally, the top 20 event pairs are selected as the salient storyline interactions for each complex topic, and the top 8 event pairs are selected for each simple topic. Table 3 shows the gold standard statistics.

5.3 Parameters and Baseline Settings

The parameters in our method are set as follows. In the probabilistic mixture model, we empirically set the subtopic number $K = 5$ for each simple topic and $K = 15$ for each complex topic, which is sufficient to cover the storylines in the gold standard. λ_B is set as 0.5 to effectively absorb domain stop words. In the time continuity factor, the size of the Gaussian window is one week (7 days), which is the news cycle in journalism. By definition, the decay rate is equal to half of the window size, i.e., $\sigma = 3.5$. In coherence score calculation, μ is used to balance the influence of the entity features and the subtopic features, and we test different values of μ to investigate the importance of different features. In the coherence graph, ϵ is the user acceptance threshold of news coherence, and we test different values of ϵ to determine its best value. In Algorithm 1, the maximum number of storyline interactions $N = 20$ for each complex topic and $N = 8$ for each simple topic, to compare with the event pairs in Table 3.

Three baseline methods discussed in Section 2 were implemented for comparison with our approach. Baseline 1 (Thread) adopts the best-performing method as tested in [19]. Documents are clustered into events by agglomerative clustering, in which the similarity score of two documents d_i and d_j is defined as:

$$S_{i,j} = e^{\frac{-|\Delta t|}{T}}(0.9\, Sim(d_i, d_j) + 0.1\, Entity(d_i, d_j)), \quad (17)$$

where $\Delta t = t_i - t_j$, $Sim(d_i, d_j)$ is the cosine similarity of the documents' TF-IDF vectors, $Entity(d_i, d_j)$ is 1 if d_i and d_j share any named entities, otherwise it is 0. The clustering process is stopped once the similarity of any event pair is below a given threshold. Those events containing only one document are discarded.

To determine the strength of event dependency, the similarity of two events $\mathcal{E}_i$ and $\mathcal{E}_j$ is defined as:

$$Sim(\mathcal{E}_i, \mathcal{E}_j) = \frac{\sum_{d_i \in \mathcal{E}_i} \sum_{d_j \in \mathcal{E}_j} S_{i,j}}{|\mathcal{E}_i| \times |\mathcal{E}_j|}, \quad (18)$$

where $|\mathcal{E}_i|$ is the number of documents in $\mathcal{E}_i$. Higher values of $Sim(\mathcal{E}_i, \mathcal{E}_j)$ indicate stronger event dependency. For each complex topic, we select the top 20 event pairs with the strongest dependency (i.e., $Sim(\mathcal{E}_i, \mathcal{E}_j)$) as the results. For each simple topic, we select the top 8 event pairs with the strongest dependency as the results. All parameters take the same values as in [19], and the centroid document of each cluster is used to illustrate the corresponding event.

Baseline 2 (Evolution)[20] splits the data into time intervals, each spans two weeks and is half overlapping with the previous one. In each interval, the probabilistic mixture model in Subsection 4.1.1 is used to extract K subtopics from the partitioned data. The evolution distance of two subtopics z_i and z_j from different intervals (suppose z_i is from an earlier interval) is determined by their KL-divergence as:

$$KL(z_j||z_i) = \sum_{w \in \mathcal{W}_{ij}} p(w|z_j) \log \frac{p(w|z_j)}{p(w|z_i)}, \quad (19)$$

where $\mathcal{W}_{ij}$ is the vocabulary of the two intervals. For each complex topic, $K = 15$ and we select the top 20 subtopic pairs with the shortest evolution distances (i.e., $KL(z_j||z_i)$) as the results. For each simple topic, $K = 5$ and we select the top 8 subtopic pairs with the shortest distances as the results. The document with the highest probability $p(z_i|d)$ in each subtopic z_i is used to illustrate the corresponding event.

Table 3. Gold Standard Statistics

Topic	$\#\mathcal{S}$	Storyline Cases	Event Pairs
BP Oil Spill*	12	Well sealing, Environment, Investigation, Compensation, etc.	Top 20
Euro Debt Crisis*	15	Greek government, EU rescue, Protests, Germany, etc.	Top 20
Haiti Earthquake*	8	Casualties, Relief efforts, International responses, etc.	Top 20
Arab Spring*	12	H. Mubarak, Protests, Military, Court trials, Media, etc.	Top 20
Iceland Ash*	8	Volcano, Environment, Air traffic, Public criticism, etc.	Top 20
Islands Clash 2012*	8	Chinese government, Japanese government, Protests, etc.	Top 20
Thailand Riot†	3	Thai government, Red-Shirts, Military	Top 8
Chilean Miner†	2	Trapped miners, Rescue efforts	Top 8
Islands Clash 2010†	2	Chinese government, Japanese government	Top 8
Russian Fire†	3	Wildfires, Russian government, Relief efforts	Top 8

Note: $\#\mathcal{S}$ is the number of storylines in a news topic.

Baseline 3 (Transformation)[21] first identifies the actors in a news topic. The words with the highest probabilities $p(w|z_k)$ for each subtopic z_k in the mixture model are extracted as the actors. Then, the proximity of two documents d_i and d_j is defined as:

$$P_{i,j} = e^{\frac{-|\Delta t|}{T}} \times Sim(d_i, d_j) \times Jaccard(K_i, K_j), \quad (20)$$

where $\Delta t = t_i - t_j$. $Sim(d_i, d_j)$ is the cosine similarity of the documents' TF-IDF vectors. K_i contains the actors in d_i, and $Jaccard(K_i, K_j)$ is the ratio of the actors shared by d_i and d_j.

The documents are clustered by the proximity measurement, and then the importance scores of actor transformations in each cluster are calculated. Five types of actor transformation are investigated, including create, continue, cease, merge, and split. The importance score of each type of transformation is determined by the term frequency and the co-occurrence frequency of actors in the cluster, which is similar to the classical interpretation of PMI (see (5)). The strength of document correlation is equal to the average importance score of all transformations in the document pair. For each complex topic, we select the top 20 document pairs with the strongest correlations from all clusters as the results. For each simple topic, we select the top 8 document pairs with the strongest correlations as the results. All parameters take the same values as in [21].

5.4 Evaluation Metric

Given a news topic, the document pairs selected by a tested method are compared with the event pairs in the gold standard (i.e., the salient storyline interactions) to evaluate the algorithm performance. The evaluations conducted in the work adopted as our baselines[19-21] all involve human judgments by reading each document to determine if it is narrating the desired event in the gold standard. However, even simple manual evaluation on a large scale over a few quality metrics would require very expensive human efforts, which is difficult to conduct on a frequent basis.

In this paper, we propose an automatic storyline interaction evaluation framework by measuring the content similarity between the selected documents and the labeled events. When matching a document pair (d_i, d_j) selected by a tested method to an event pair ($\mathcal{E}_i$, $\mathcal{E}_j$) in the gold standard, a key problem is to judge whether a document d_i is narrating the desired event $\mathcal{E}_i$. In the gold standard, each event $\mathcal{E}_i$ is depicted by a paragraph excerpt $\mathcal{P}_i$ in the corresponding Wikipedia article. However, a news article d_i cannot be directly compared with a paragraph excerpt $\mathcal{P}_i$ for two reasons. Firstly, the length of a news article is much longer than that of a Wikipedia paragraph excerpt. In our dataset, the average length of a news article is 612 words, which is much longer than the average length of a paragraph excerpt (55 words) in the gold standard. Secondly, Wikipedia, as the largest example of participatory journalism, is collaboratively crafted by netizens of various professions in different countries, thus its writing style and terminology is different from the news articles written by professional journalists[31].

Instead, we use the news articles cited by a Wikipedia paragraph excerpt $\mathcal{P}_i$ to represent the event $\mathcal{E}_i$. In the gold standard, on average a paragraph excerpt is linked to 1.78 news articles as its reference, and at least one news article is cited by any paragraph excerpt. The average length of a cited news article is 754 words, which is close to the average length of a news article (612 words) in the dataset. Therefore, we compare a document d_i with the news articles cited by the paragraph excerpt $\mathcal{P}_i$ to match the event $\mathcal{E}_i$.

ROUGE[32] is used to measure the content similarity between the news articles, which counts the number of overlapping units (e.g., n-gram) between the candidate document and a set of reference documents. Formally, ROUGE is defined as:

$$ROUGE\text{-}N = \frac{\sum_{d \in \{References\}} \sum_{gram_n \in d} Count_{\text{match}}(gram_n)}{\sum_{d \in \{References\}} \sum_{gram_n \in d} Count(gram_n)}, \quad (21)$$

where $gram_n$ represents an n-gram instance, and $Count_{\text{match}}(gram_n)$ is the accumulated number of n-grams co-occurred in the candidate document and the reference documents *References*.

ROUGE is recall-oriented, and a candidate document in which n-grams are shared by multiple references is favored by the metric. This is reasonable since there could be multiple good news reports on an important event, and an article which is more similar to consensus among reference documents is more likely to cover the whole event well. A previous study[34] shows that automatic evaluation using the unigram version of ROUGE (i.e., ROUGE-1) correlates well with human judgment. In addition, the word-level evaluation in ROUGE-1 (whereas $gram_n$ in (21) represents a word, including named entity) fits into the word entropy analysis in our methodology. Therefore, ROUGE-1 is adopted to measure the content similarity between the candidate document and the reference documents in this paper.

The problem of matching a document pair (d_i, d_j) selected by a tested method (suppose $t_{d_i} < t_{d_j}$)[12] to an event pair $(\mathcal{E}_i, \mathcal{E}_j)$ in the gold standard (suppose $t_{\mathcal{E}_i} < t_{\mathcal{E}_j}$)[13], is transformed into matching the two document-event pairs $\{d_i, \mathcal{E}_i\}$ and $\{d_j, \mathcal{E}_j\}$ respectively. When matching a candidate document d_i to a labeled event $\mathcal{E}_i$, the time difference between d_i and $\mathcal{E}_i$ is firstly checked. Only if $|t_{d_i} - t_{\mathcal{E}_i}| \leqslant 1$ can make d_i and $\mathcal{E}_i$ a possible match. The ROUGE-1 score of d_i and the reference documents of $\mathcal{E}_i$ is then calculated, after stop words removal and stemming[34]. If the ROUGE-1 score exceeds a certain threshold δ, the candidate document d_i is judged to match the desired event $\mathcal{E}_i$.

To determine the threshold δ, we manually selected 100 matched document-event pairs, and calculated their ROUGE-1 scores. The average ROUGE-1 score of matched pairs is 0.340, and the minimum ROUGE-1 score of sampled matched pairs is 0.278. We then randomly generate 10 000 unmatched document-event pairs. The average ROUGE-1 score of unmatched pairs is only 0.115, and the maximum ROUGE-1 score of unmatched pairs is below 0.2. Thus, we set the threshold as the median value of the minimum matched ROUGE-1 score and the maximum unmatched ROUGE-1 score, i.e., $\delta = 0.24$.

Both of the document-event pairs (i.e., $\{d_i, \mathcal{E}_i\}$ and $\{d_j, \mathcal{E}_j\}$) are required to be matched for a matched event pair (i.e., a valid storyline interaction). For each news topic, a tested method outputs the same number of document pairs as in the gold standard, and each document pair is mapped to each of the event pairs in the gold standard to check if it is a valid storyline interaction. The accuracy of finding the valid storyline interactions in a news topic is used to evaluate the algorithm performance, which is defined as:

$$Accuracy = Count_{\text{match}}(N)/N, \tag{22}$$

where $N = 20$ for each complex topic and $N = 8$ for each simple topic. $Count_{\text{match}}(N)$ is the number of matched event pairs (i.e., valid storyline interactions) in a news topic.

5.5 Experimental Results

We evaluate the proposed algorithm by three sets of experiments: 1) algorithm parameter analysis, 2) baseline method comparison, and 3) user preference incorporation.

5.5.1 Algorithm Parameter Analysis

In the first set of experiments, we analyze the influence of the parameter values on the performance of the proposed algorithm, with no user preference given.

As discussed in Subsection 5.3, ϵ is the threshold of coherence that readers can accept. Once the coherence score exceeds ϵ, the two documents are connected in the coherence graph. We determine the value of ϵ in the following manner. For each news topic, we randomly sample at most 100 document pairs in each month from the corpus. Then, the coherence score of each sampled document pair is calculated, and all the sampled document pairs are ranked by their coherence scores. Finally, we set the value of ϵ to be the coherence score of the document pair which is ranked 5%, 25%, 50%, 75%, and 100% of all sampled pairs respectively. All the other parameters in the algorithm are set as in Subsection 5.3, in which the entity relatedness factor is determined by the topic-level NPMI, and $\mu = 0.5$ in coherence score calculation. Fig.5 shows the mean accuracy on the 10 topics in our dataset.

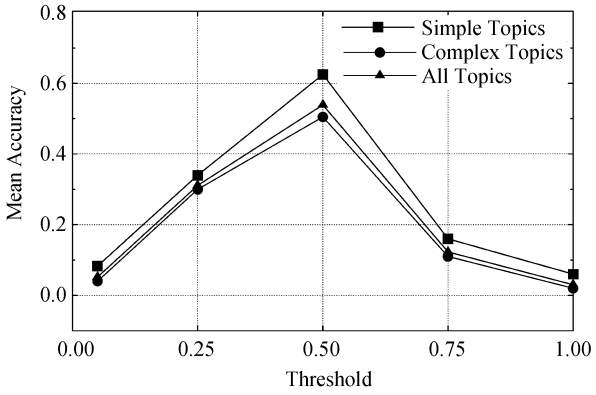


Fig.5. Results of coherence threshold.

The results in Fig.5 show that for both complex and simple topics, a moderate value of ϵ is desired to produce the best result (e.g., 50% ranking). If the threshold is too high (e.g., 5% ranking), only those document pairs with the highest coherence scores can be connected. Since similarity implies coherence in our definition, a document pair with similar content is favored by the coherence factors. Thus, most of the coherent document pairs (w.r.t the high threshold) come from the same storylines, while those document pairs in which the two documents are in different storylines are not connected. As a result, most of the storyline interactions cannot be captured. If the threshold is too low (e.g., 75% or 100% ranking), any pair of document with little correlation can be connected, and the coherence graph G is close to a complete graph (in each sliding window). Thus, the scores of the vertices assigned by PageRank are indistinguishable from each other. The informative events are randomly selected, and the method fails.

[12] If d_i and d_j are published on the same date, then we compare their time stamps in minutes.
[13] The time of an event is the date labeled in the paragraph excerpt in the corresponding Wikipedia article.

In coherence score calculation, μ is used to balance the influence of the entity features and the subtopic features. To investigate the importance of the two types of features in our method, we set the value of μ to be 0, 0.25, 0.5, 0.75, and 1 respectively. All the other parameters in the algorithm are set as in Subsection 5.3, in which the entity relatedness factor is determined by the topic-level NPMI, and ϵ is set as the 50% ranking score of all sampled pairs. Fig.6 shows the mean accuracy on the 10 topics in our dataset.

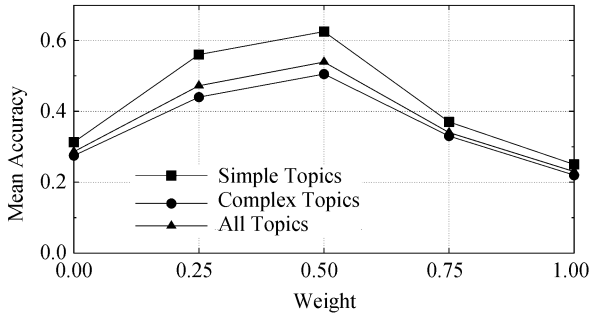


Fig.6. Results of coherence weight.

Fig.6 shows that the entity features and the subtopic features are complementary to each other, and combining both types of features achieves better result than using any of them alone. If only subtopic features are considered (i.e., $\mu = 1$), the document pairs in which the two documents are in different storylines will receive lower coherence scores. Thus, many storyline interactions will be missed. If only entity features are considered (i.e., $\mu = 0$), a large part of information in the news articles (e.g., What and Why) is not considered, and two documents sharing a few entities are connected even though they have little correlation. This will add much noise to the coherence graph, and the method cannot effectively discover the salient storyline interactions.

Thirdly, we compare different methods for the entity relatedness factor, i.e., Jaccard, word-level NPMI, and topic-level NPMI. Details of the three methods are explained in Subsection 4.1.2. μ is set to 0, thus only the entity relatedness factor is considered. All the other parameters in the algorithm are set as in Subsection 5.3, and ϵ is the 50% ranking score of all sampled pairs. Fig.7 shows the mean accuracy on the 10 topics in our dataset, and the results validate our analysis in Subsection 4.1.2. Compared with Jaccard and word-level NPMI that directly match the entity words, the proposed topic-level NPMI can effectively measure the entity relatedness in a news topic.

Finally, the algorithm performs better for simple topics than for complex topics. Simple topics have fewer storylines with simpler story development than

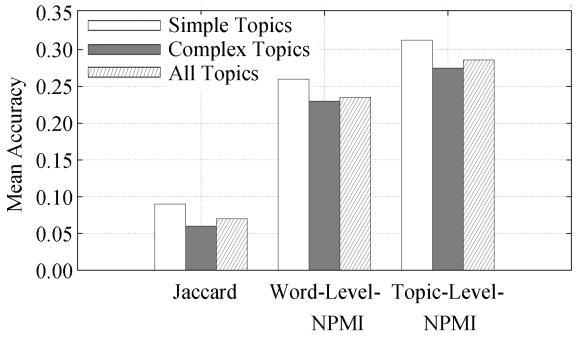


Fig.7. Results of entity relatedness.

topics. The storyline interactions in a simple topic are more distinguishable from each other and from other events, thus they are easier to be identified.

5.5.2 Baseline Method Comparison

In the second set of experiments, we compare the proposed algorithm (with the best parameter setting tested in Subsection 5.5.1) against the three baseline methods discussed in Subsection 5.3, with no user preference given. Fig.8 shows the mean accuracy on the 10 topics in our dataset.

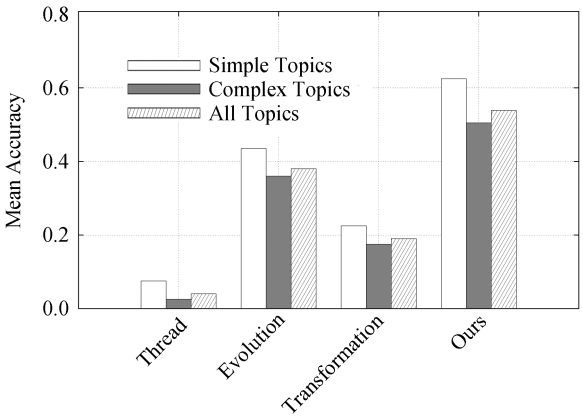


Fig.8. Results of baseline comparison.

The results in Fig.8 show that the proposed method outperforms the baselines on both simple and complex topics. Baseline 1 (Thread) performs poorly as event dependency is totally determined by document similarity, thus the discovered document pairs depict the development of single storylines, not the interactions of different storylines. Baseline 3 (Transformation) suffers from similar problem as it narrows the search space based on document proximity, thus misses many storyline interactions before the actor transformation analysis. Baseline 2 (Evolution) utilizes topic consistency in different intervals to link coherent document pairs, and the method has a relatively favorable performance. However, the method splits the corpus of different top-

ics into fixed-length time intervals, and only documents from different intervals can be connected. Thus, those storyline interactions inside an interval cannot be discovered. In addition, the method does not consider entity relatedness in news coherence, which also limits its effects.

Fig.2 shows two excerpts of the structured overview of "Islands Clash 2012*" (Fig.2(a)) and "Islands Clash2010†" (Fig.2(b)). In Fig.2(a), compared with the topic's timeline by Wikipedia, the generated overview can identify key events with their possible causes and consequences, and reveal the mutual influence among different storylines. For example, both the activists detention (Aug. 15th) and the official visit (Aug. 18th) in *Japanese government* triggered the first wave of anti-Japanese protests (since Aug. 19th); the protest restriction (Sept. 16th) in *Chinese government* terminates the development of the storyline *Chinese protests* (after Sept. 19th); and the second wave of anti-Japanese protests (since Sept. 13th) is the *Eye of the Storm* of the news topic. In contrast, the overview in Fig.2(b) is much simpler, with only two storylines involved. Although the two topics share the same background (i.e., disputes over the same islands), the clash in 2010 is more local and is much less influential than the clash in 2012, in which violent protests occurred and the economy was highly impacted. From the storyline interactions illustrated in the overview, readers can acquire a clear picture of the news topic, and understand the influence of key events on the story development.

Finally, although the proposed algorithm outperforms the three baseline methods in our dataset, there are still some limitations of the work that should be concerned and be improved in future. For example, the entity relatedness factor used to measure the coherence is relied on the accuracy of named entity identification in news articles. The Stanford NER tools we used for entity extraction has a mean accuracy of 87.67% (Person 91.88%, Organization 82.91%, Location 88.21%, as reported in [25]), which means around 12% identified entities may be incorrect. To mitigate this problem, the proposed method averages the effects of all entity pairs to calculate the entity relatedness factor, as shown in (9). In future, we plan to measure the salience of entities with regard to a news topic, and only select those salient entities to calculate the entity relatedness factor.

Furthermore, this paper assumes that a news event is focused on a single aspect of the story, thus it can only belong to one storyline in a topic. Under this assumption, the interactions of storylines are represented as coherent event pairs, instead of solo events. If an event is related to multiple aspects of the story, the proposed method selects the most relevant storyline the event is focused on, and then investigates its interactions with other storylines. To better illustrate various storyline developments in real news browsing scenarios, we plan to give a more flexible definition of the storyline, and enhance the current model to support multiple representations of storyline interactions.

5.5.3 User Preference Incorporation

In the third set of experiments, we demonstrate the flexibility of the proposed algorithm for user preference, compared with traditional keyword search. Two news topics are selected for our case study: "BP Oil Spill" from complex topics and "Chilean Miner" from simple topics. For each news topic, two queries are issued, each representing a storyline interesting to a reader. The baseline method searches the query in the topic's timeline by Wikipedia, and returns event summaries that best match the query words. Our method integrates the user query into the PageRank algorithm, and discover the informative events biased towards the desired storylines. Table 4 shows the settings of the case study. Results of each method (i.e., event summaries or informative events) are manually checked by the domain experts to evaluate the algorithm's performance.

Table 4. User Preference Case Study

Topic	Preference	User Query
BP Oil Spill*	Well sealing	"leak, well, seal, effort"
	Compensation	"BP, compensation, victim, claims"
Chilean Miner†	Trapped miners	"trap, health, shelter, food"
	Rescue efforts	"rescue, effort, drill, plan"

In the BP case, compared with the search results of the baseline method, the query-specific results of our algorithm capture more events in the desired storyline, and are more favored by the domain experts. The reason is that many events in a complex topic are very complicated, and cannot be well depicted by a few keywords as in the user query. The baseline method performs extremely poorly when querying for the storyline *Compensation*, since the query words "compensation", "victim", and "claims" are not found in the timeline summary by Wikipedia, and event summaries retrieved by the query word "BP" come from various storylines in the topic. We have observed that some event summaries in the timeline are indeed in the desired storyline, such as "Obama meets with Svanberg, Hayward, McKay. BP agrees to fund a $20 billion escrow account." (Jun. 16th). Those summaries are written without the query words, thus cannot be retrieved by keyword search. On the contrary, our method successfully captures the major events in the storyline, in which the query words are integrated into the topic-sensitive vectors, and relevant

events are scored higher than the events in other storylines. Thus, our algorithm can better understand the user interest than keyword search for a complex news topic.

In the Miner case, however, our algorithm does not show significant advantages over the baseline method. For each query, both methods can successfully retrieve major events in the desired storyline. The reason is that for a simple topic, most of the events can be well depicted by a few keywords, and different storylines can be effectively differentiated by the issued query. For example, most of the event summaries in the storyline *Rescue efforts* by Wikipedia contain the query words, and well cover the relevant events, such as "Second collapse hampers *rescue efforts* and blocks access to lower parts of the mine. *Rescuers* begin *drilling* boreholes to send down listening devices." (Aug. 7th); "First attempt to *drill* a hole to *rescue* the men, *Plan* A, begins." (Aug. 30th).

Finally, we conduct a user survey to quantitatively compare the performance of the proposed method and the baseline method in user preference incorporation. Similar to the settings in the user preference case study, for each of the ten news topics in our dataset, a query is issued representing a storyline interesting to a reader. The output results (i.e., event summaries by the baseline method or informative events by our method) are read by three domain experts, each of whom will assign a score of 1 to 5 according to his/her satisfaction of the results. A rank of 5 (or 1) indicates that the results of the method is the most (or least) satisfying for the given user preference. Table 5 shows the average ratings given by the domain experts on the ten news topics.

Table 5. User Preference Quantitative Measurement

Topic	Preference	Baseline	Ours
BP Spill*	Compensation	1.0	**3.0**
Euro Debt*	EU rescue	2.3	**3.0**
Haiti Quake*	Relief efforts	2.7	**3.3**
Arab Spring*	Military	2.0	**3.0**
Iceland Ash*	Public criticism	2.3	**3.3**
Diaoyu 2012*	Chinese government	3.3	**3.7**
Thai Riot†	Red-Shirts	4.0	**4.3**
Chile Miner†	Rescue efforts	**4.3**	**4.3**
Diaoyu 2010†	Chinese government	**5.0**	4.7
Russian Fire†	Russian government	**4.7**	**4.7**

The results in Table 5 are consistent with the observations we have made in the user preference case study. For a complex news topic, we conclude that the proposed method that projects the user query into the subtopic space can achieve higher user's satisfaction than traditional keyword search. For a simple news topic, although keyword search can obtain favorable results in meeting the user's need, it still lacks the information of event (and storyline) relationship, which can be illustrated by the proposed method.

6 Conclusions

In this paper, we studied a novel text mining problem of exploring the storyline interactions in a news topic, which can help news readers navigate the story development in a global as well as a focused view. The proposed approach addresses the coherence between news articles, and can more effectively discover salient storyline interactions than traditional similarity-based methods. User preference can be naturally integrated into our method to generate query-specific results, which outperforms keyword search in the timeline summary.

Future work will collect more data and test our approach on a larger variety of news topics. We also plan to optimize the algorithm by jointly modeling the coherence and importance of news articles, and develop more reliable metrics to evaluate the algorithm's performance. Finally, we will extend this research from online news to other fields, such as the academic interactions in scientific literatures.

Acknowledgment We thank the anonymous reviewers for their valuable comments.

References

[1] Kovach B, Rosenstiel T. Blur: How to Know What's True in the Age of Information Overload. New York: Bloomsbury, 2011.

[2] Hunziker H. In the Eye of the Reader: Foveal and Peripheral Perception. Germany: Staubli, 2007.

[3] Allan J, Carbonell J, Doddington G *et al.* Topic detection and tracking pilot study: Final report. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Feb. 1998, pp.194-218.

[4] Yan R, Wan X, Otterbacher J *et al.* Evolutionary timeline summarization: A balanced optimization framework via iterative substitution. In *Proc. the 34th SIGIR*, Jul. 2011, pp.745-754.

[5] Yan R, Kong L, Huang C, Wan X, Li X, Zhang Y. Timeline generation through evolutionary trans-temporal summarization. In *Proc. EMNLP*, Jul. 2011, pp.433-443.

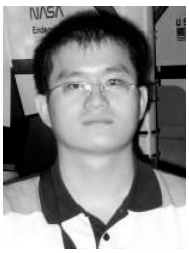
[6] Hu P, Huang M, Xu P, Li W, Usadi A K, Zhu X. Generating breakpoint-based timeline overview for news topic retrospection. In *Proc. ICDM*, Dec. 2011, pp.260-269.

[7] Wang D, Li T, Ogihara M. Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs. In *Proc. AAAI*, Jul. 2012, pp.683–689.

[8] Pustejovsky J, Castano J, Ingria R *et al.* TimeML: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering 2003*, Maybury M T (ed.), 2003, pp.28-34.

[9] Seker S E, Diri B. TimeML and Turkish temporal logic. In *Proc. ICAI*, Jul. 2010, pp.881-887.

[10] Reichenbach H. Elements of Symbolic Logic. London: Macmillan Co., 1947.
[11] Rosu G, Bensalem S. Allen linear (interval) temporal logic — Translation to LTL and monitor synthesis. In *Proc. the 18th CAV*, Aug. 2006, pp.263-277.
[12] Ahmed A, Xing E P. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proc. UAI*, Jul. 2010, pp.20-29.
[13] Jo Y, Hopcroft J E, Lagoze C. The web of topics: Discovering the topology of topic evolution in a corpus. In *Proc. the 20th WWW*, Apr. 2011, pp.257-266.
[14] Cui W, Liu S, Tan L *et al.* TextFlow: Towards better understanding of evolving topics in text. *IEEE Trans. Visualization and Computer Graphics*, 2011, 17(12): 2412-2421.
[15] Gao Z J, Song Y, Liu S *et al.* Tracking and connecting topics via incremental hierarchical Dirichlet processes. In *Proc. the 11th ICDM*, Dec. 2011, pp.1056-1061.
[16] Lin C X, Mei Q, Han J, Jiang Y, Danilevsky M. The joint inference of topic diffusion and evolution in social communities. In *Proc. the 11th ICDM*, Dec. 2011, pp.378-387.
[17] Wang Y, Agichtein E, Benzi M. TM-LDA: Efficient online modeling of latent topic transitions in social media. In *Proc. the 18th SIGKDD*, Aug. 2012, pp.123-131.
[18] Teh H W, Jordan M I, Beal M J, Blei D M. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Proc. Neural Information Processing Systems*, Dec. 2005, pp.1385-1392.
[19] Nallapati R, Feng A, Peng F, Allan J. Event threading within news topics. In *Proc. the 13th CIKM*, Nov. 2004, pp.446-453.
[20] Mei Q, Zhai C. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proc. SIGKDD*, Aug. 2005, pp.198-207.
[21] Choudhary R, Mehta S, Bagchi A, Balakrishnan R. Towards characterization of actor evolution and interactions in news corpora. In *Proc. the 30th ECIR*, Mar. 2008, pp.422-429.
[22] Shahaf D, Guestrin C, Horvitz E. Trains of thought: Generating information maps. In *Proc. the 21st WWW*, Apr. 2012, pp.899-908.
[23] Ansell G. Introduction to Journalism (2nd edition). Johannesburg: Jacana Media (Pty) Ltd., 2011.
[24] Bowles H. Storytelling and Drama: Exploring Narrative Episodes in Plays (8th edition). Amsterdam: John Benjamins Publishing Company, 2010.
[25] Finkel J R, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proc. ACL*, Jun. 2005, pp.363-370.
[26] Hofmann T. Probabilistic latent semantic indexing. In *Proc. the 22nd SIGIR*, Aug. 1999, pp.50-57.
[27] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 1977, 39(1): 1-38.
[28] Endres D M, Schindelin J E. A new metric for probability distributions. *IEEE Trans. Information Theory*, 2003, 49(7): 1858-1860.
[29] Thagard P, Verbeurgt K. Coherence as constraint satisfaction. *Cognitive Science*, 1998, 22(1): 1-24.
[30] Haveliwala T H. Topic-sensitive PageRank. In *Proc. the 11th WWW*, May 2002, pp.517-526.
[31] Lih A. Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proc. the 5th International Symposium on Online Journalism*, Apr. 2004, pp.1-31.
[32] Lin C Y. ROUGE: A package for automatic evaluation of summaries. In *Proc. ACL Workshop on Text Summarization Branches Out*, Jul. 2004, pp.74-81.
[33] Lin C Y, Hovy E H. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. NAACL*, May 2003, pp.71-78.
[34] Porter M F. An algorithm for suffix stripping. *Program*, 1980, 14(3): 130-137.



Po Hu is a fifth year Ph.D. student in the Department of Computer Science and Technology, Tsinghua University, Beijing. He received his bachelor degree from the Department of Computer Science and Technology, Tsinghua University, in 2009. Po has published and co-authored several research papers in ICDM, CIKM, and SIGKDD conferences. His research interest includes textual temporal analysis, document retrieval, and topic modeling.

Min-Lie Huang is an associate professor in the Department of Computer Science and Technology, Tsinghua University. He received his Ph.D. degree in computer science and technology from the Department of Computer Science and Technology, Tsinghua University in 2006. He is the associate editor of BMC Bioinformatics, and PC member of ACL 2012, EMNLP 2012, EACL 2012, NA-ACL 2012, and CIKM 2012. He has published and co-authored 15 research papers in ACL, COLING, IJCAI, ICDM and AAAI conferences. His research interest includes natural language processing, data mining, and machine learning.

Xiao-Yan Zhu is a professor and Ph.D. supervisor of Department of Computer Science and Technology, Tsinghua University. She received her bachelor degree from the University of Science and Technology of Beijing in 1982, and received her Ph.D. degree in computer science and technology from the Nagoya Institute of Technology, Japan, in 1990. Prof. Zhu is the director of the State Key Laboratory of Intelligent Technology and Systems and member of CCF. She has published and co-authored over 50 research papers in ACL, COLING, IJCAI, SIGKDD, ICDM, CIKM, and AAAI conferences, and KAIS, JCST, and BMC Bioinformatics journals. Her research interest includes pattern recognition, neural networks, machine learning, natural language processing, and data mining.