Summarizing Complex Events: a Cross-modal Solution of Storylines Extraction and Reconstruction

Shize Xu
xsz@pku.edu.cn

Shanshan Wang
cheers_echo_mch@163.com

Yan Zhang[*]
zhy@cis.pku.edu.cn

Department of Machine Intelligence, Peking University, Beijing, China
Key Laboratory on Machine Perception, Ministry of Education, Beijing, China

Abstract

The rapid development of Web2.0 leads to significant information redundancy. Especially for a complex news event, it is difficult to understand its general idea within a single coherent picture. A complex event often contains branches, intertwining narratives and side news which are all called storylines. In this paper, we propose a novel solution to tackle the challenging problem of storylines extraction and reconstruction. Specifically, we first investigate two requisite properties of an ideal storyline. Then a unified algorithm is devised to extract all effective storylines by optimizing these properties at the same time. Finally, we reconstruct all extracted lines and generate the high-quality story map. Experiments on real-world datasets show that our method is quite efficient and highly competitive, which can bring about quicker, clearer and deeper comprehension to readers.

1 Introduction

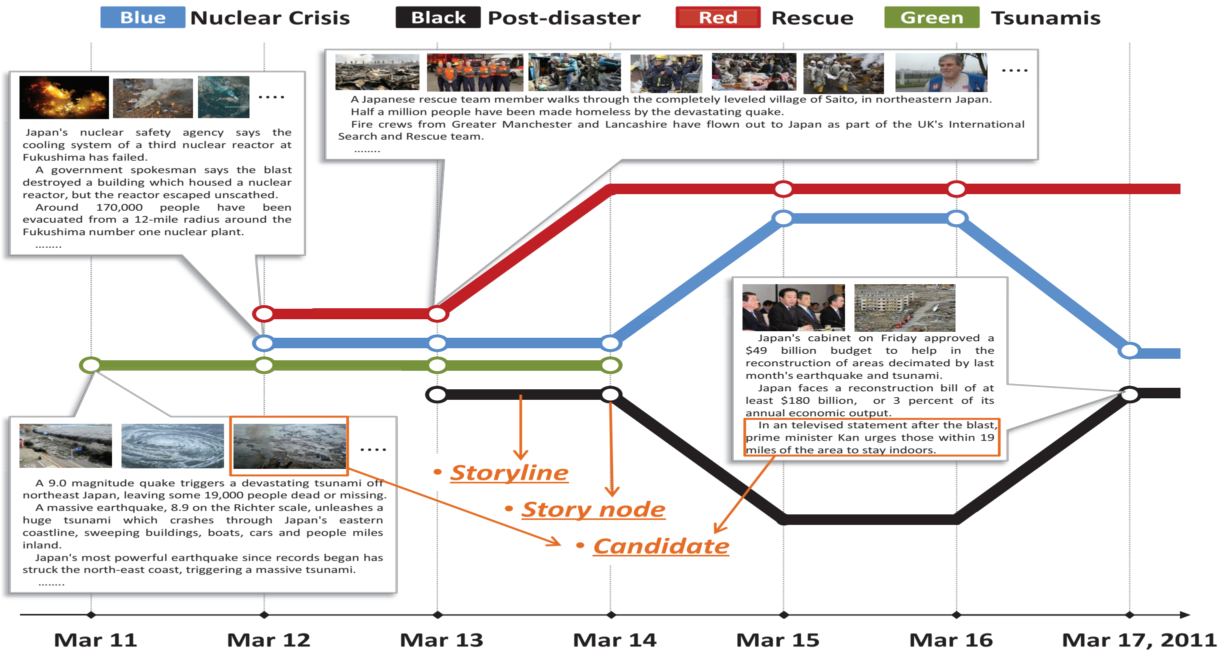
News reports usually consist of various modalities of tremendous information, especially all kinds of textual information and visual information, which make web users dazzled and lost. The situation gets worse on complex news events. To help readers quickly grasp the general information of the news, a more concise and convenient system over multi-modality information should be provided. For example, given a large collection of texts and images related to a specified news event (e.g., *East Japan Earthquake*), such a system should present a terse and brief summarization about the event by showing different clues of its development, and thus helping readers to effectively find out "when, where, what, how and why" at a glance.

The researches (Goldstein et al., 2000) on automatic multi-document summarization (MDS) have helped a lot when we generate a description for a specific event. However, it traditionally exhibits in a very simple style like a "0-dimensional" point. The appearance of *Timeline* (Allan et al., 2001) brings about a visual progress for massive documents analyses. Readers can not only get the most important ideas, but also browse the story evolution in chronological order. Previous news summarization systems with structured output (Yan et al., 2011) have focused on timeline generation. Timeline becomes a "1-dimensional" line. This style of summarization only works for simple stories, which are linear in nature. However, the structure of complex stories usually turns out to be non-linear. These stories branch into storylines, dead ends, intertwining narratives and side news. To explore these lines, we need a map to reorganize all the information. Therefore, a "2-dimensional" story map is in bad need. Figure 1 shows a part of the story map generated by our system for representing *East Japan Earthquake*. We notice that the whole event evolves into 4 branches. Each of them focuses on a specific sub-topic and is distinct from other lines. Figure 2 takes a close look at the 4 nodes from different lines, and they differ a lot from each other as expected.

Text information is more precise and exquisite when compared with images. Nevertheless, as the

[*] Corresponding author

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1281–1291,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics



Japan's nuclear safety agency says the cooling system of a third nuclear reactor at Fukushima has failed.
A government spokesman says the blast destroyed a building which housed a nuclear reactor, but the reactor escaped unscathed.
Around 170,000 people have been evacuated from a 12-mile radius around the Fukushima number one nuclear plant.
....

A Japanese rescue team member walks through the completely leveled village of Saito, in northeastern Japan.
Half a million people have been made homeless by the devastating quake.
Fire crews from Greater Manchester and Lancashire have flown out to Japan as part of the UK's International Search and Rescue team.
....

Japan's cabinet on Friday approved a $49 billion budget to help in the reconstruction of areas decimated by last month's earthquake and tsunami.
Japan faces a reconstruction bill of at least $180 billion, or 3 percent of its annual economic output.
In an televised statement after the blast, prime minister Kan urges those within 19 miles of the area to stay indoors.

A 9.0 magnitude quake triggers a devastating tsunami off northeast Japan, leaving some 19,000 people dead or missing.
A massive earthquake, 8.9 on the Richter scale, unleashes a huge tsunami which crashes through Japan's eastern coastline, sweeping buildings, boats, cars and people miles inland.
Japan's most powerful earthquake since records began has struck the north-east coast, triggering a massive tsunami.
....

Figure 1: Four storylines are obtained in the story map of "East Japan Earthquake". They focus on *Tsunamis*, *Nuclear Crisis*, *Rescue* and *Post-disaster* respectively.

saying goes, "a picture paints a thousand words", an image could provide far more information than words do. In fact, a summarization including both texts and images will absolutely yield a more powerful and intuitive description about the news event. Under this motivation, we study on extracting and reconstructing tracks with different sub-topics for a complex event. To the best of our knowledge, the exploration and analysis of 2-dimensional cross-modal summarization is academically novel.

We are faced with two main problems. The first is how to select the most important sentences and images to make up the final story map. The previous work by Shahaf et al. presents a 2-D story map called "metro map", which summarizes the complex topics (Shahaf et al., 2012). They study on the document-level, and use the entire news document as one story node. But on real web, this may confront some difficulties. On one hand, news articles may report the event from different perspectives, especially those reviews or retrospective reports. This kind of documents contains many useful saliency information of different sub-topics, but they cannot be further subdivided to help understand each better. On the other hand, some documents, such as cover news and interviewing reports, contain one or two famous remarks. Since these documents also include too much useless information, it's inappropriate to use the whole document as a story node. So our work, the sentence-level story map extraction, is just aimed at this brand new problem. The second problem is the way to utilize the cross-modal

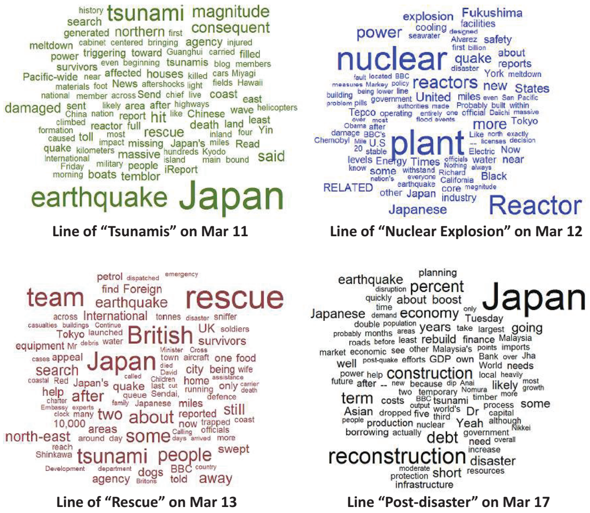


Figure 2: Word distributions vary a lot among nodes in different storylines

information suitably. Our goal is not only to fuse sentences and images together, but also to provide a unified framework to improve them mutually.

In this paper, we introduce a novel solution for the story map summarization problem. All the sentences and images are the candidates for making up the final map. The analysis of complicated information usually requires a semantic-level knowledge study. We address this in the pre-processing of data in Section 3.1. The key task of our research is the extraction of storylines. We reveal two fundamental properties of an ideal storyline, and propose an optimization algorithm in Section 3.2. A highly compatible MDS sub-algorithm is also fused in and offers help to the sentences/images selection. In Section 3.3, the extracted storylines are reconstructed as a final story map. The experimental results conducted on four real datasets show that our approach can perform effectively.

The rest of this paper is organized as follows. Some related researches are demonstrated in Section 2. We introduce our methodology in Section 3. The experimental results in Section 4 prove the effectiveness of our approach. Finally, we conclude this paper and present our future work in Section 5.

2 Related Work

Generally speaking, multi-document summarization can be either extractive or abstractive. Researchers mainly focus on the former which extracts the information deemed most important to the summary. Various techniques have been used for this type of MDS (Haghighi and Vanderwende, 2009; Contractor et al., 2012). Graph-based text summarization techniques have been widely used for years. The algorithms, used in TextRank (Mihalcea and Tarau, 2005) and LexPageRank (Radev et al., 2004), which are meant to compute sentence importance, are similar to those in PageRank and HITS.

Recently, timeline becomes a popular style to present a schedule of events and attracts many researchers consequently. For example, Yan et al. make use of timestamps to generate an evolutionary timeline (Yan et al., 2011). Shahaf et al. present a 2-D story map called "metro map" to summarize the complex topic (Shahaf et al., 2012). But previous work only studies on the document level, which inevitably brings about much information redundancy.

Previous studies show that the use of visual materials not only leads to the conservation of information but also promotes comprehension (Panjwani et al., 2009). Thus the cross-modal fusion is necessary. Wu et al. propose a framework of multimodal information fusion for multimedia data analysis by learning the optimal combination of multimodal information with the superkernel fusion (Wu et al., 2004). Borrowing the idea of recommendation in heterogeneous network into the cross-modal news summarization is also a convincing research. Xu et al. tackle this task and bring out an 1-D cross-media timeline generation framework (Xu et al., 2013).

Summarization of multimedia involves researches on information retrieval of multimedia. Since textual and visual information are quite different from each other, how to make a good transformation to mine latent knowledge from unannotated images is of great concern. Feng and Lapata (2010) use visual words to describe visual features and then propose a probabilistic model based on the assumption that images and their co-occurring textual data are generated by mixtures of latent topics.

However, to the best of our knowledge, no existing research manages to generate a 2-dimensional story map automatically and integrate images and texts into a unified framework at the same time.

3 Methodology

The original data of one event is a collection of news documents on different days, or in a finer granularity, a set of sentences and images with different timestamps. Each item in the data collection is a candidate for selection to form the final map. We denote the data collection as C, and $C = C_s \cup C_v$, where C_s is the subset containing all sentences and C_v contains all images. In the following elaboration of our method, dateset C is the important knowledge base. As the ultimate goal for a specific event, we would like to generate the "2-D" story map $\mathcal{M}$, whose main component is a set of "1-D" storylines $\mathcal{L} = \{L_1, L_2, \ldots\}$. Each storyline $L \in \mathcal{L}$ is made up of a set of "0-D" story nodes, $L = \{I_1, I_2, \ldots\}$, each of which is composed of a set of candidates sharing the same timestamp $I = \{c_1, c_2, \ldots\}$. For more concise, our method can be scheduled with the following three steps:

1. Prepare semantic knowledge for each candidate, and then purify data collection C by eliminating the noisy candidates;
2. Conduct OPT-LSH algorithm to extract $\mathcal{L}$ that contains all qualified storylines;
3. Reconstruct the storylines as the final map $\mathcal{M}$.

3.1 Pre-processing

Before we extract the storylines, we have to solve two problems first. Since our work is cross-modal, the semantic knowledge under the literal and visual surface is basically required. Recall from Section 2, there exist many effective ways to dig into the semantic level of image and text. In this paper, we employ the approach proposed by Jiang and Tan (2006). They present a convincing approach which employs a multilingual retrieval model to apply knowledge mining on semantic level. The first is the sentence feature vector generation. With the preprocessing such as stemming and stop words removal, they extract the textual TF-IDF feature Vec_s of each sentence. The second, also the challenging part, is the image feature vector generation. In this step, for each region they extract visual features that are consisting of 6 color features and 60 gabor features which have been proven to be useful in many applications. Color features are the means and variances of the RGB color spaces. Gabor features are extracted by calculating the means and variations of the filtered image regions on 6 orientations. After the visual feature vectors of the image regions are extracted, all image regions are clustered using the k-means algorithm. The generated clusters, called "visterms" or "visual words", are treated as a vocabulary for the images. Besides visual features, they also utilize the context textual feature of each image as the semantic supplement to generate to final feature vector Vec_v.

Based on these feature vector of each cantidate, they further calculating the intra-modal similarity with classical IR methods, they obtain the inter-modal similarity through the vague transformation (Mandl T, 1998). We also note that translation tools such as VIPS (Cai et al., 2003) and WebKit[1] can help us to segment web documents and pick those text blocks whose coordinates are neighboring to each specified image. In this way we can successfully obtain images, text contexts and content sentences. Three kinds of semantic similarity are now ready. They are uniformly denoted as $sim(c_i, c_j)$, representing the similarity between two candidates. c_i and c_j can be any type of modalities.

Another problem is the noise from irregular data. We would like to utilize the sentences and images of high quality. The intuitive assumption is that a good candidate should have substance in speech, and be coherent with other good candidates. Fortunately, many useful measures are now available. They can be used to choose better candidates. Inspired by the analysis analogy to information retrieval, we extend the idea of the classical *PageRank* algorithm to estimate the authority for each candidate. The similarity between two candidates is regarded as the weighted "link" between them.

Inspired by the idea of classic "update summarization" task, we try to avoid those chronologically ordered documents sets focusing on a constant topic. Therefore, given the particularity of our task, we also have to develop the weighting function Γ with the temporal factor before starting the ranking algorithm. Our fundamental assumption is that the inter-date and inter-modal "links" have different influence comparing with the intra ones. Therefore the core formula calculating the authority of c_i is adapted as follows to make it become weight-compatible:

$$Auth(c_i) = \frac{1-q}{|C|} + q \cdot [\ \alpha \sum_{c_j \in C'} \Gamma(c_i, c_j) \cdot \frac{Auth(c_j)}{O(c_j)} \\ + (1-\alpha) \sum_{c_k \in C''} \Gamma(c_i, c_k) \cdot \frac{Auth(c_k)}{O(c_k)}\ ]$$

C' is the subset of C which only includes the candidates of the same modality with c_i, and C'' is the cross-modal candidates subset. $O(c_j)$ denotes the out-degree of c_j, and smoothing parameter q is set as the common value 0.85. Parameter α is used to balance the biases of intra- and inter-modal impact. If α is set to 1, it means the cross-modal information is abandoned, and vice versa. $\Gamma(c_i, c_j)$ contains two terms as follows:

$$\Gamma(c_i, c_j) = sim(c_i, c_j) \cdot e^{\frac{-\|c_i.t - c_j.t\|}{2\sigma^2}}$$

[1] http://www.webkit.org

The second term of Γ's formula is Gaussian Kernel (Aliguliyev R, 2009), which is used to measure the temporal gap between two candidates ($*.t$ denotes the date-based timestamp). Note that the similarity metric is content-based and time-independent, since the time decay function is only used to adjust the ranking impact strength. In this way, we can give higher authority to the more informative and coherent candidates. The optimal value of σ, which controls the spread of kernel curves, is sensitive to datasets and will be discussed later.

We eliminate the candidates whose authority goes under threshold. The data collection C is then significantly downsized and purified for later work. Authority is also used to determine the presentation sequence inside each story node of final map.

3.2 Extraction: LSH-OPT Algorithm

Other traditional relative methods need to pre-decide how many sub-topics are going to be obtained, like clustering or other supervised models. They fail in the unsupervised automatic storylines extraction problem. In this Section, we propose a concrete algorithm to extract storyline set $\mathcal{L}$ from C. Our proposed LSH-based algorithm can automatically optimize the number of storylines according to their self-evaluation.

3.2.1 Task Formalization

Let's investigate the storyline first. We may think of some basic attributes as well as many extension properties. The number of nodes in the storyline L (denoted as $|L|$) is intuitively one of its basic attributes. We would like to define another basic attribute called the SUPPORT of L. In detail, $Support(L) = \min_{I_k \in L} |I_k|$, which denotes the smallest size among all the story nodes it has.

The most challenging part is to properly model extension properties. We observe that an effective storyline should meet three key requirements: (1) **Coherence**. Within one storyline, news changes gradually as time goes and the evolution indicates consistency among component story nodes. We rely on the notion of coherence developed in *Connect-the-Dots* (Shahaf and Guestrin, 2010) and transform it to what we exactly need in this research; (2) **Diversity**. According to MMR principle (Goldstein et al., 1999), though the work is about summary, we still can draw an analogy and derive that a good storyline should be concise and contain redundant information as few as possible, i.e., two sentences providing information of similar content should not be presented in different storylines; (3) **Coverage**. The extracted storyline set $\mathcal{L}$ should keep alignment with the source collection C, which is intuitive and even proved to be significant as proposed in (Li et al., 2009). However, *Coverage* in some ways is technically redundant in front of *Diversity*. We decide to use the first two criteria in extraction process and use the last one to verify the effectiveness.

Since a storyline is composed of several nodes, we can select or abandon nodes mainly according to these two requirements. In fact, both of them involve a measurement of similarity between two story nodes, denoted by two word distributions (see Figure 2). Specifically, for story node I_i, its distribution probability of word w is estimated as $p(w|I_i) = \frac{\sum_{c \in I_i} TF(w)}{\sum_w \sum_{c \in I_i} TF(w)}$ where the denominator is used for normalization. Then *Kullback-Leibler* divergence is employed to denote the distance between two nodes I_i and I_j:

$$D_{KL}(I_i, I_j) = \sum_w p(w|I_i) \log \frac{p(w|I_i)}{p(w|I_j)}$$

In addition, we introduce the decreasing and increasing variants based on logistic functions, $D^\flat_{KL} = 1/(1 + e^{D_{KL}})$ and $D^\sharp_{KL} = e^{D_{KL}}/(1 + e^{D_{KL}})$, to map the distance into $[0, 1]$. Given the measurement, we can formulate the two properties.

For *Coherence*, a storyline L_i consists of a series of individual but correlated nodes, which do not necessarily have the serial timestamps. We would like to choose such a set of nodes $\{I_1, I_2, \ldots\}$, and at the same time guarantee this criterion:

$$Cor(L_i) = \frac{1}{|L_i|} \sum\nolimits_{1 \le k < |L_i|} D^\flat_{KL}(I_k, I_{k+1})$$

For *Diversity*, each storyline $Li \in \mathcal{L}$ should demonstrate quite different subtopics with other storylines. This is the most essential motivation for us to step into 2-dimensional field. This criterion can be used to maximize the minimum diversity value among all storylines:

$$Div(\mathcal{L}) = \min_{L_i, L_j \in \mathcal{L}} \{ \frac{\sum_{I_k \in L_i} \sum_{I_{k'} \in L_j} D^\sharp_{KL}(I_k, I_{k'})}{|L_i| \cdot |L_j|} \}$$

Then the problem can be transformed into the following optimization problem. Parameters θ_1, θ_2 and θ_3 denote the minimum number of nodes in each line, the smallest size of candidates in each node and the coherence lower bound respectively. The task is to extract an optimal $\mathcal{L}$ out of C, such that:

$$\begin{aligned} \forall L \in \mathcal{L},\ |L| \geq \theta_1 \,\&\, Support(L) \geq \theta_2; \\ \forall L \in \mathcal{L},\ Cor(L) \geq \theta_3; \\ Div(\mathcal{L})\ is\ maximized. \end{aligned}$$

3.2.2 Optimization Algorithm

It can be proved that finding the optimal set $\mathcal{L}$ is an NP-Complete problem (not presented due to the limited space). Thus the brute-force exhaustive approach is crashed. We develop a near-optimal algorithm based on locality sensitive hashing (OPT-LSH). The original LSH solution is a popular technique used to solve the nearest neighbor search problems in high dimensions. Its basic idea is to hash similar input items into the same bucket (i.e., uniquely definable hash signature) with high probability. All potential storylines can be targeted fast if we make good use of this idea.

LSH performs probabilistic dimension reduction of high dimensional data by projecting a higher d-dimensional vector Vec_c (recall from Section 3.1) to a lower d'-dimensional vector ($d' << d$), such that the candidates which are in close proximity in the higher dimension get mapped into the same item in the lower dimensional space with high probability. It guarantees a lower bound on the probability that two similar input items fall into the same bucket in the projected space and also the upper bound on the probability that two dissimilar vectors fall into the same bucket (Indyk and Motwani, 1998).

One of the key requirements for good performance of LSH is the careful selection of the family of hashing functions. In OPT-LSH, we use the hashing scheme proposed by Charikar (Charikar M, 2002). In detail, d' random unit d-dimensional vectors $\vec{r_1}, \vec{r_2}, \ldots, \vec{r_{d'}}$ are generated first. Each of the d entries of $\vec{r_i}$ is drawn from a standardized normal distribution N(0,1). Then the d' hashing functions are defined as:

$$h_i(c) = \begin{cases} 1, & if\ \vec{r_i} \cdot Vec(c) \geq 0 \\ 0, & if\ \vec{r_i} \cdot Vec(c) < 0 \end{cases} \quad 1 \leq i \leq d'$$

We represent the d'-dimensional bucket feature $\mathbf{h}$ for c, $\mathbf{h}(c) := [h_1(c), \ldots, h_{d'}(c)]$. There are $2^{d'}$ different buckets at most. Each denotes a potential storyline, so we have to verify the probability of similar candidates falling into the same bucket, whose lower bound is given by Charikar (Charikar M, 2002).

Simply filtering and searching among all potential lines in single pass may lead to empty result set if in post-processing no bucket satisfies all constraints. This could probably happen because the input parameter d' is set so large that the optimal set of candidates is separated into different buckets. However, we will get a suboptimal result in turn when d' is too small. We are then motivated to tune LSH by iterative relaxation that varies d' in each iteration. Changing the value of d' balances the leverage between expected number of potential storylines and their properties. We perform a binary search between 1 and d' to identify the ideal number of hash functions to employ. Algorithm 1 shows the pseudo code of our OPT-LSH algorithm. The LSH time is bounded by $O(d'|C| \log |C|)$ since the binary search relaxation iteration runs for $\log |C|$ times in the worst case and the hashing time is $O(d'|C|)$.

3.3 Storylines Reconstruction

At last, we manage to reconstruct all the storylines in $\mathcal{L}^{opt}$. A real-world storyline may sometimes intertwine with another, educe other branches, and end its own evolvement. The way to reconstruct a more effective layout of the story map requires further study and provides a good research direction in the future. However, in this paper we order the sentences/images in each story node according to their authority scores. Next, all storylines are arranged to proceed along the timestamps, thus a storyline never turns back in the map. Then we adjust the structure to make story nodes sharing the same timestamp stay close, though they belong to different lines. Figure 1 shows the sample output of our system.

4 Experiments

4.1 Dataset

There is no existing standard evaluation data set for 2-dimensional cross-modal summarization methods. We randomly choose 4 news topics from 4 selected news websites: *New York Times*, *BBC*, *CNN* and

Algorithm 1 OPT-LSH Algorithm

```
Input: Candidate set $C$, similarity function $sim$,
    bucket dimensions $d'$
Output: A near-optimal storylines set $\mathcal{L}^{opt}$
  // Main Algorithm
  Initialize $left = 1, right = d', max = -1$
  repeat
    $d' = (left + right)/2$
    $\mathcal{L} \leftarrow LSH(C, d')$
    Revise all word distributions $p(w|I)$ in $\mathcal{L}$
    if $\forall L \in \mathcal{L}$, $|L| \geq \theta_1$ and $Support(L) \geq \theta_2$ and
    $Cor(L) \geq \delta_3$ then
      if $Div(\mathcal{L}) > max$ then
        $\mathcal{L}^{opt} \leftarrow \mathcal{L}, max \leftarrow Div(\mathcal{L})$
      end if
      $left = d' + 1$
    else
      $right = d' - 1$
    end if
  until $left > right$
  return $\mathcal{L}^{opt}$

  // $LSH(C, d') : \ Buckets$
  Generate $d'$ unit vectors randomly
  for $j = 1$ to $|C|$ do
    for $i = 1$ to $d'$ do
      if $\vec{r_i} \cdot Vec(c_j) \geq 0$ then
        $h_i(c_j) \leftarrow 1$
      else
        $h_i(c_j) \leftarrow 0$
      end if
    end for
    $\mathbf{h}(c_j) = [h_1(c_j), \ldots, h_{d'}(c_j)]$
  end for
  return $Buckets \leftarrow \{\mathbf{h}(c_j) | c_j \in C\}$
```

Reuters. We query each event confined to these sites and crawl webpages' html docs. Referring to Section 3.1, timestamps, text contents, images and their text contexts are extracted. Table 1 shows the details. These 4 datasets all contain massive information and complex evolutions.

4.2 Analysis of Our System

Table 1: Statistics of Datasets.

Event (Query)	Document	Time Span	σ	α
EJE	504	Mar 11-Apr 8, 2011	3	0.9
OWS	638	Sept 17-Dec 10, 2011	12	0.7
NBA	489	July 1-Dec 28, 2011	17	0.6
ME	437	June 21-Aug 31, 2011	9	0.7

* Abbreviations EJE, OWS, NBA, ME denote *East Japan Earthquake*, *Occupy Wall Street*, *NBA Lockout* and *Murdock's Eavesdropping* respectively.

Since there is no standard for us to verify the effectiveness of our solution, we have to utilize convincing criteria based on manual evaluation of experts, and then we can compare them with other approaches. In order to setup the system, based on the optimization problem shown in Section 3.2, we assign the value of 1 to both θ_1 and θ_2, and empirically set θ_3 as 0.6 which can balance the number of potential lines and the quality of story map as well. Then the problem can be re-interpreted in natural language as follows. Given the data collection, we manage to find out a set of storylines such that every line in it contains at least one non-null story node and keeps self-coherence not less than 0.6. What's more, the diversity of the whole set is maximized. Before comparing OPT-LSH with other systems, we do some further analysis of inherent properties first.

4.2.1 Compactness

The essential idea of summarization is to reduce the data size, so that a more concise representation will be generated and help users to fast grasp the main points. Therefore the *Compactness* of a story map needs to be guaranteed. In the pre-processing module, we have already excluded significant number of inferior candidates with the extended PageRank. Nevertheless what we really care is the compactness that OPT-LSH brings about. In the candidates and story nodes selection processes, only the most saliency and coherent candidates can appear in the final representation. We count the number of sentences and images in $\mathcal{L}^{opt}$, denoted as $||\mathcal{L}^{opt}||$, and then we compare it with the collection size $|C|$ (both before and after pre-processing) to test the compactness. Table 2 shows that OPT-LSH further reduces the representation scale significantly.

4.2.2 Coverage

Obviously, only the verification of compactness is far from enough. As mentioned in Section 3.2, the storyline set $\mathcal{L}$ we extract should keep alignment with the source collection, and contain informative as well as comprehensive information in C. Thus we

Table 2: The compactness of OPT-LSH

Dataset	EJE	OWS	NBA	ME
$\|C\|$-before	12049	22890	20403	10237
$\|C\|$-after	1454	2357	1187	1042
$\|\|\mathcal{L}\|\|$	87	168	113	92
Downsizing	94.0%	92.9%	90.5%	91.2%

need to verify another property of our summarization, the *Coverage*. Inspired by (Shannon C, 2001), we employ the *Information Entropy* to represent the information quantity based on solid mathematical theory. The less information quantity decreases after summarizing, the more story comprehensiveness is maintained. In this way we verify the property of coverage. Particularly, *Shannon* denotes the entropy H as follows of a discrete random variable X, which in fact is a word distribution. The base knowledge in our work is the global probability mass function $P(W_C)$ based on the entire vocabulary of C, with possible values $\{p(w_1), \ldots, p(w_{|W_C|})\}$.

$$\begin{aligned} H(X) &= E\{-\log(P(W_C))\|X\} \\ &= -\sum_{w_i \in X} p(w_i)\log(p(w_i)) \end{aligned}$$

Although different word distributions may have the same H, we do not focus on the similarity of two corpora, but the difference of information quantities they are carrying. So H is an ideal criterion.

Besides the comparison of the entropies of C and $\mathcal{L}$, it gives us a chance to study different modules' contributions in our solution. There are two places that we may simplify the solution. One is the feature of temporal gap. If we set the parameter σ to $infinity$, then we can remove the second term of the calculation of Γ (i.e. $\sigma \leftarrow \infty$) and then bring out a *time-insensitive* system. The other is the cross-modal feature. We set parameter α as 1 to make the system work in one single modality, and ignore all images (only the textual sentences are available) to make up story map. The work then becomes the study with *text-bias*. We also implement the *simplest* system that blocks images as well as temporal feature. Figure 3 highlights the good performance of our system. We are maintaining information by a larger proportion of the original data collection. Considering the property of high compactness, our solution tackles the information redundancy quite well, and promisingly delivers entire knowledge with compact structure.

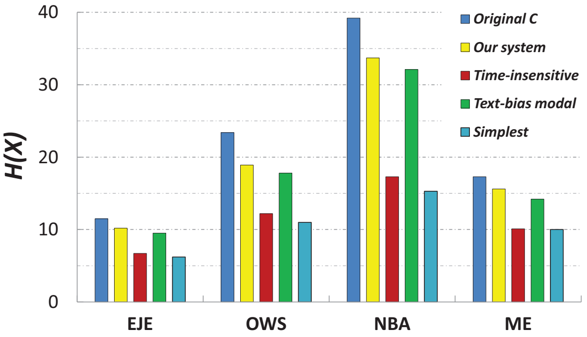


Figure 3: The y-axis denotes the entropy. And the larger H is, the richer information it brings.

During our experiments on *Coverage*, we have some interesting findings. Datasets perform differently when we take different values of σ and α, which controls the temporal decaying rate and cross-modal learning respectively. The events with short life-cycles prefer a smaller value of σ to dominate the influence from neighbors, as well as the intra-modal bias. On the contrary, long-living events prefer lager σ and more inter-modal bias to get information replenishment from different dates and modality. Due to the limited space we don't present the tuning details, but the optimal values are shown in Table 1. In fact, using the cross-modal mutual influence can, more or less, help to improve the effectiveness of information extraction and summarization.

4.3 User Study

Before we introduce other existing methods that can also tackle the cross-modal 2-dimensional summarization problem, we have to setup the appropriate standards to quantify users' evaluation.

4.3.1 Metrics

In the user study, we evaluate the effectiveness of our story maps in aiding users to integrate different aspects of multi-faceted information. (Shahaf et al., 2012) also focuses on story map generation and puts forward two convincing metrics to answer the following questions:

- *Micro-Knowledge*: Can the maps help users retrieve information faster than other methods?
- *Macro-Knowledge*: Can the maps help users understand the big picture better than other methods?

For *micro-knowledge*, we wish to see how maps help users answer specific questions. We compare the level of knowledge attained by users using our method with two other systems: *Google News* and *TDT*. Google News is a computer-generated site that aggregates headlines from news sources worldwide. News-viewing tools are dominated by portal and search approaches, and *Google News* is a typical representative of those tools. *TDT* (Nallapati et al., 2004) is a successful system which captures the rich structure and dependencies of news events.

We have noticed that making comparisons between different systems is not convincing, since the output of Google News and TDT is different both in content and in presentation (and in particular, cannot be double-blind). In order to isolate the effects of sentence selection vs. map organization, we introduce a hybrid system into the study: the system with *structureless* story map displays the same sentences and images as our system but with none of the structure. Its output is basically the same with our full system but with a single storyline and merges node content for each date. And each story nodes are sorted chronologically and displayed similarly to Google News. We implement TDT based on (Nallapati et al., 2004) (cos + TD + SimpleThresholding), and pick a representative article from each cluster. The purpose of the study is to test a single query. We also obtain the results from Google News using the same queries.

We recruit 64 volunteers to browse all the four events, and one of the four systems is assigned to each person randomly. After browsing, users are asked to answer a short questionnaire (8 questions), composed by domain experts. Users answer as many questions as possible in limited time (8 minutes). The statistics of their answers are promised to evaluate the micro-knowledge on different systems. In order to aid in comprehension, we give some examples about those asked questions. For the event of EJE, we ask that

1. *How many magnitude was initially reported by the USGS, and what about the finally report?*
2. *List at least six countries that had dispatched their rescue teams.*
3. ...

And for OWS, we ask that

Table 3: *Macro-knowledge* performance on four datasets

Dataset	Our System	Google News	TDT
EJE	56.3%	23.2%	20.5%
OWS	62.2%	22.1%	15.7%
NBA	58.3%	18.9%	22.8%
ME	47.2%	26.3%	26.5%

1. *What was the attitude of President Obama about the protesters on October?*
2. *When did the protesters begin dressing "corporate zombies" in New York?*
3. ...

These questions can effectively help us to investigate users' $micro\text{-}knowledge$ about the events.

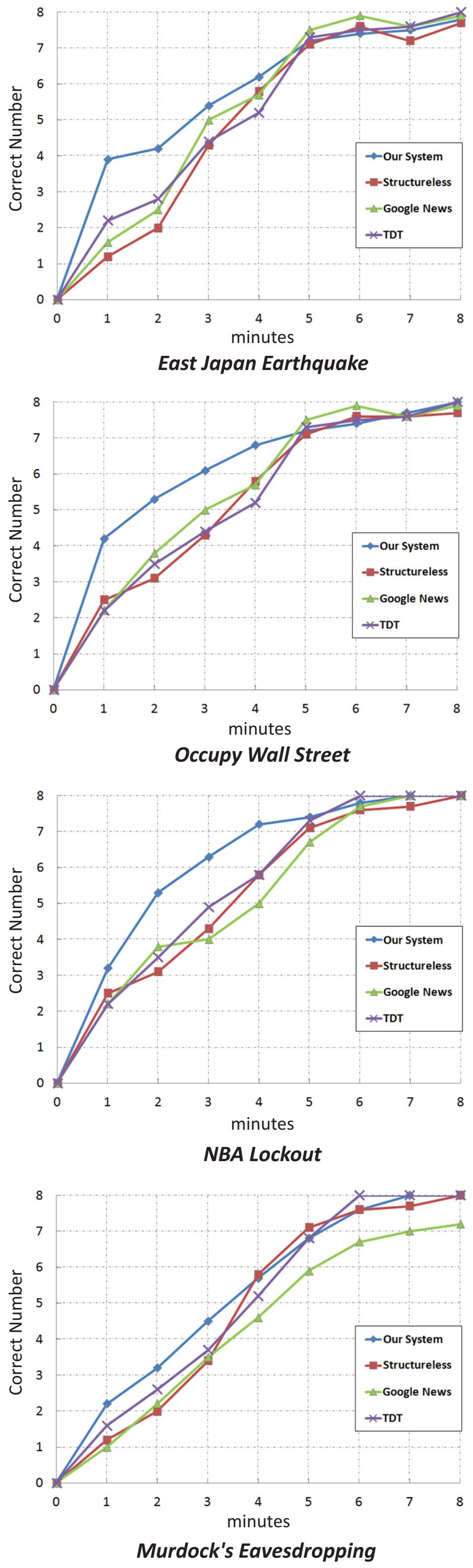
As for *macro-knowledge*, unlike the retrieval study that evaluates users' ability to answer questions, we are interested in the use of story maps as high-level overviews, allowing users to understand the big picture. We believe that the ability to explain a certain issue is the only proof of understanding. Therefore, the 64 volunteers are then asked to write four paragraphs to summarize the four events respectively. This time, all three systems' (the structureless system presents the same content as our system) results are provided and we let users choose the sentences with complete freedom. Then we count the number of sentences they employed from each system and derive the average proportions. According to the results we can research on the macro-knowledge that different systems deliver.

4.3.2 Results

We take the time cost and the average numbers of correct answers of different systems to evaluate on the micro-knowledge. Figure 4 shows the results.

We can find out that our system outperforms the others significantly when users taking less time to learn the knowledge. The failure of structureless system proves that our work of storyline reconstruction makes lots of advantages to help reading.

On the other hand, Table 3 analyzes the statistics of macro-knowledge. It's also obvious that users would like to refer to the sentences that our system provides. A reasonable explanation is that the story maps we generate can clarify users' thoughts and views on the complicated events.



East Japan Earthquake

Occupy Wall Street

NBA Lockout

Murdock's Eavesdropping

Figure 4: *Micro-knowledge* performance on four datasets

Table 4: Average runtime of different datasets

Dataset	EJE	OWS	NBA	ME
$\|C\|$	1454	2357	1187	1042
Iterations	3.5	4.7	5.4	3.7
Runtime (ms)	1582	3214	3089	1314

4.4 Runtime Analysis

At last, we analyze the time performance of our OPT-LSH algorithm on a PC server (16G RAM, 2.67GHz 4-processors CPU). The average iterations for different initial value of d' and the runtime are shown in Table 4. The results are acceptable.

5 Conclusions

In this paper, we study the feasibility of automatically generating cross-modal story maps and present a novel solution to this challenging problem. Our works mainly tackle the problems of storylines extraction and reconstruction. Specifically, we investigate two requisite properties of an ideal storyline, *Coherence* and *Diversity*. Then the convincing criteria are devised to model both. We formalize the task as an optimization problem and design an algorithm to solve it. Classical IR and text analyzing techniques like PageRank are fused into the unified framework, and a near-optimal solution is employed to deal with the NP-complete problem. Experiments on web datasets show that our method is quite efficient and competitive. We also verify that it brings quicker, clearer and deeper comprehension to users.

As a future work, we plan to adapt parameters automatically on the basis of different types of datasets. Improving the layout quality of story map by concerning the interactivity of different media (e.g. images order) is also significant. Furthermore, our framework is universal, so that the media other than text and image can be adopted as well.

Acknowledgments

We sincerely thank all the anonymous reviewers for their valuable comments, which have helped to improve this paper greatly.

This work is supported by NSFC with Grant No. 61073081 and 61370054, and 973 Program with Grant No. 2014CB340405.

References

Agrawal R, Gollapudi S, Kannan A, et al. 2011 *Enriching textbooks with images*. Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, pages 1847-1856.

Aliguliyev R M. 2009 *A new sentence similarity measure and sentence based extractive technique for automatic text summarization*. Expert Systems with Applications, 2009, 36(4): 7764-7772.

Allan J, Gupta R, Khandelwal V. 2001 *Temporal summaries of new topics*. Proceedings of the 24th Annual International ACM Conference on SIGIR, pages 10-18.

Cai D, Yu S, Wen J R, et al. 2003 *VIPS: a visionbased page segmentation algorithm*. Microsoft Technical Report, MSR-TR-2003-79.

Charikar M S. 2002 *Similarity estimation techniques from rounding algorithms*. Proceedings of the 34th Annual ACM Symposium on Theory of Computing, ACM, pages 380-388.

Chen Y, Jin O, Xue G R, et al. 2010 *Visual contextual advertising: Bringing textual advertisements to images*. Proceedings of the 24th AAAI Conference, AAAI, pages 1314-1320.

Contractor D, Guo Y, Korhonen A. 2012. *Using Argumentative Zones for Extractive Summarization of Scientific Articles*. COLING, pages 663-678.

Evans D K, McKeown K, Klavans J L. 2005 *Similarity-based multilingual multi-document summarization*. IEEE Transactions on Information Theory, pages 1858C1860.

Feng Y, Lapata M. 2010 *Topic models for image annotation and text illustration*. Human Language Technologies: The 2010 Annual Conference of NAACL, pages 831-839.

Goldstein J, Mittal V, Carbonell J, Kantrowitz M. 2000 *Multi-document summarization by sentence extraction*. Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4. Association for Computational Linguistics, pages 40-48.

Goldstein J, Kantrowitz M, Mittal V, et al. 1999 *Summarizing text documents: sentence selection and evaluation metrics*. Proceedings of the 22nd Annual International ACM Conference on SIGIR, ACM, pages 121-128.

Haghighi A, Vanderwende L. 2009. *Exploring content models for multi-document summarization*. Proceedings of Human Language Technologies: The 2009 Annual Conference of NAACL, Association for Computational Linguistics, pages 362-370.

Indyk P, Motwani R. 1998 *Approximate nearest neighbors: towards removing the curse of dimensionality*. Proceedings of the 30th Annual ACM Symposium on Theory of Computing, ACM, pages 604-613.

Jiang T, Tan A H. 2006 *Discovering image-text associations for cross-media web information fusion*. Knowledge Discovery in Databases: PKDD 2006, pages 561-568.

Li L, Zhou K, Xue G R, et al. 2009 *Enhancing diversity, coverage and balance for summarization through structure learning*. Proceedings of the 18th International Conference on World Wide Web, ACM, pages 71-80.

Mandl T. 1998 *Vague transformations in information retrieval*. ISI 1998, pages 312-325.

Mihalcea R, Tarau P. 2005 *A language independent algorithm for single and multiple document summarization*. Proceedings of IJCNLP 2005.

Nallapati R, Feng A, Peng F, et al. 2004 *Event threading within news topics*. Proceedings of the 13rd ACM International Conference on Information and Knowledge Management, ACM, pages 446-453.

Panjwani S, Micallef L, Fenech K, et al. 2009 *Effects of integrating digital visual materials with textbook scans in the classroom*. International Journal of Education and Development using ICT, 2009, 5(3).

Radev D R, Jing H, et al. 2004 *Centroid-based summarization of multiple documents*. Information Processing & Management, 2004, 40(6): 919-938.

Radev D, Winkel A, Topper M. 2002 *Multi document centroid-based text summarization*. ACL Demo Session, 2002.

Shahaf D, Guestrin C. 2010 *Connecting the dots between news articles*. Proceedings of the 16th ACM Conference on SIGKDD, ACM, pages 623-632.

Shahaf D, Guestrin C, Horvitz E. 2012 *Trains of thought: Generating information maps*. Proceedings of the 21st International Conference on World Wide Web, ACM, pages 899-908.

Shannon C E. 2001 *A mathematical theory of communication*. ACM SIGMOBILE Mobile Computing and Communications Review, 2001, 5(1): 3-55.

Wu Y, Chang E Y, Chang K C C, et al. 2004 *Optimal multimodal fusion for multimedia data analysis*. Proceedings of the 12th annual ACM International Conference on Multimedia, ACM, 2004: 572-579.

Xu S, Kong L, Zhang Y. 2013 *A cross-media evolutionary timeline generation framework based on iterative recommendation*. Proceedings of the 3rd ACM conference on International Conference on Multimedia Retrieval, ACM, pages 73-80.

Yan R, Wan X, Otterbacher J, et al. 2011 *Evolutionary timeline summarization: a balanced optimization framework via iterative substitution*. Proceedings of the 34th International ACM Conference on SIGIR, ACM, pages 745-754.