2007

Individualized Storyline-based News Topic Retrospection

Fu-ren Lin
National Tsing Hua University, frlin@mx.nthu.edu.tw

Feng-mei Huang
National Tsing Hua University, jalumay@gmail.com

Chia-hao Liang
Institute for Information Industry, chliang@iii.org.tw

144. Individualized Storyline-based News Topic Retrospection

Fu-ren Lin
Institute of Technology Management
National Tsing Hua University
frlin@mx.nthu.edu.tw

Feng-mei Huang
Institute of Technology Management
National Tsing Hua University
jalumay@gmail.com

Chia-hao Liang
IDEAS
Institute for Information Industry
chliang@iii.org.tw

Abstract

It takes a great effort for common news readers to track events promptly, and not to mention that they can retrospect them precisely after it occurred for a long time period. Although topic detection and tracking techniques have been developed to promptly identify and keep track of similar events in a topic and monitor their progress, the cognitive load remains for a reader to digest these reports. A storyline-based summarization may facilitate readers to recall occurred events in a topic by extracting informative sentences of news reports to compose a concise summary with essential episodes. This paper proposes SToRe (Story-line based Topic Retrospection), that identifies events from news reports and composes a storyline summary to portray the event evolution in a topic. It consists of three main functions: event identification, main storyline construction and storyline-based summarization. The main storyline guides the extraction of representative sentences from news articles to summarize occurred events. This study demonstrates that different topic term sets result in different storylines, and in turn, different summaries. This adaptation is useful for users to review occurred news topics in different storylines.

Keywords: topic retrospection, clustering, text mining, event threading, summarization

Introduction

The prevalence of Internet technologies diversifies the information aggregation and dissemination. Internet users have new ways to receive information via Internet, such as e-mails, electronic news, blogs, etc. Information regarding events occurring around the world can be known via these channels. People may track the development of news topics for individual or institutional interests. For example, for a company in an industry, it is beneficial to keep eye on news reports of its partners and competitors to gain its competitive intelligence. The amount of information which people can effectively consume is limited. Many techniques have been developed to ease this cognitive load, such as search engines (information retrieval), automated categorization, clustering, and recommendation (Berghel 1997; Brown 2000; Sebastiani 2002) .

Information overloading is not only associated with the quantity of information, but also with information format and quality (Ho et al. 2001). Search engines, although fast and comprehensive, only present their results as lists of hits that a user needs to read through to differentiate relevant from irrelevant information. Search engines contribute to the dimension of information quantity. To face constituently evolving news events, topic detection and tracking (TDT) techniques have been developed to group articles corresponding to a topic and track future events belonging to that topic. Mechanisms developed in TDT research

focus on the dimension of information quality. People cannot quickly comprehend the sketch of a topic from clustered news sets. However, most researches merely present a list of news reports, and ignore the information presentation.

Existing TDT techniques and systems usually present an event in a topic as a cluster of news report using clustering techniques. For people to review events occurred long time ago, it takes a great effort to pick an anchor report in the cluster as the entry point to cognitively recognize the event. However, it is also a cognitively expensive and time consuming for readers to comprehend the event evolution of a news topic. In the study we provide a better information presentation to ease such cognitive load and facilitate readers to quickly capture the theme of a news topic. The selection of events and presentation sequence of sentences create different storylines, and in turn, generate different summaries.

A new set of techniques for recalling and assembling occurred events in a topic is necessary to sketch different storylines based on reviewers' interests and then summarize news stories. This study proposes techniques for three main tasks of topic retrospection: *event identification*, *main storyline construction* and *storyline-based summarization*. These techniques are derived from existing information retrieval and text mining, such as term extraction, clustering, and sentence extraction for summarization. Techniques from building graphs and growing maximum spanning tree are adopted to draw storylines.

A news topic catching the public attention recently in Taiwan is the deployment of electronic toll collection (ETC) system to two national freeway systems. The chaos occurred when this BOT (Build-Operate-Transfer) project was run by a private company charged excess handling fees. It also raised many political and judicial cases as this disorder became a headlight news. However, as the memory decays, people can barely recall events occurred in the turmoil. Thus, this study collects news articles on this topic as the test bed to exercise proposed techniques and to summarize this news topic in different storylines built by different anchor news articles having different frequencies of topic terms. The results show that the proposed techniques for topic retrospection can automate the summarization of a news topic with different storylines biased by reviewers' preferences via different anchor articles. Section 2 introduces the framework of topic retrospection, and Section 3 demonstrates these techniques for the news topic on ETC deployment. Section 4 analyzes the effects of different anchor news articles with different frequencies of topic terms on the storyline construction, and in turn, the summarization. Section 5 discusses findings and Section 6 concludes this study and lays out future research.

Topic Retrospection

Definitions

This study adopts the definitions of topic and event in TDT (Allan et al. 1998; Franz et al. 2001). A topic is defined as *a seminal event or activity, along with all directly related events and activities*, e.g., HP's acquisition of Compaq. An event is *something* (*non-trivial*) *that happens at a particular time and place*, e.g., the board of HP approved the acquisition on March 19, 2002. A story is a *news report on an event.* Consequently, a topic is composed of a series of related events, and can be talked as a storyline distilled from news reports. Similar to a film (topic), there are a main story and many episodes (events) to compose the plot. If an automated mechanism can quickly capture the main story and important episodes, a moviegoer can quickly overview the movie. Hence, the hierarchy of the news can be formed into a topic, events and stories.

Furthermore, in order to distinguish this research from related literatures (McKeown 2002; Radev et al. 2004), this study defines "*topic retrospection*" to differentiate it from TDT, event threading and multi-source summarization. The main distinction to TDT and event threading is that topic retrospection further filters, organizes, summarizes topic with text. Additionally, taking the topic structure to compose a summarized article is also different from the multi-source summarization. Hence, *topic retrospection is an integrated mechanism to identify various events under a news topic and construct relations among these events to summarize news articles in order to present users the sketch of event evolution.*

This study defines the research scope as follows. In this research, a set of *n* news stories $\boldsymbol{S} = \{s_1, s_2, ..., s_n\}$ belonging to a certain topic $\Im$ will be given. $\boldsymbol{S}$ is divided into *m* events $\varepsilon = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_m\}$. Besides, this study assumes that (1) each story describes one of the events in ε, (2) each story only belongs to at most one event in ε, and (3) the chronological relationship of two stories can be established if the first story of the first event is earlier than the first story of the second event.

The structure of a topic is expressed by a directed graph. Each vertex in the graph denotes an event, and an edge represents the dependency between two vertices. If the edge is established between a pair of events, the direction of edge implies that they exist in time-ordering sequence and the earlier event is more likely to influence the latter one. It is hard to approve the causality because of lacking the text semantic understanding. However, there exists a certain degree of influences according to the similarity of terms.

Topic retrospection mechanisms

The proposed topic retrospection mechanism analyzes the event structure from a collection of news reports on a topic to compose the summary which provides a news reader a quick review of the news topic. Specifically, the proposed mechanism consists of the following three main functions.

(1) *Event identification*: distinguishing various events under a news topic. The similar news stories will be clustered together to indicate an event using the clustering algorithm.
(2) *Main storyline construction*: identifying the relationship between events and how relevant these events with the main storyline.
(3) *Storyline-based summarization*: Extracting the representative sentences to compose the summary under the main theme. The summarized text, like a guideline, can also link to original reports to facilitate readers to read the news articles.

At the first stage, we adopted a SOM (self-organizing map) technique (Kohonen 1997), called GSOM (growing self-organizing map) (Dittenbach 2000) to identify different events in a topic and to form clustered event sets. A main storyline in a news topic is constructed by measuring the relevance between events and the main storyline built as a maximal spanning tree (MST). Based on the graph of topic structure, events on the main storyline and branches will be summarized into an article. The accumulated weight among different features is used to select sentences as constituent sentences to summarize the topic. Besides, the pre-process is taken prior to the aforementioned three main functions in order to convert news reports from unstructured data into the vector space model. The pre-process is summarized in Table 1. This study collects corpus from Chinese news website and segments the Chinese nature language by using CKIP (Chinese Knowledge and Information Processing) which is developed by the Institute of Information Science and the Institute of Linguistics of Academia Sinica in 1986.

Event identification

The first task in topic retrospection is to distinguish how many events happened in a topic. In previous researches (Dittenbach 2000; Shih et al. 2004), the self-organizing maps (SOM) (Kohonen 1997) has been applied in clustering different types of documents and shows its usefulness and reliability. Therefore, this study adopts the SOM for unsupervised clustering to identify events. In addition, we cannot anticipate the number of events beforehand since it is subject to change by different topics. Hence, this study adopts GSOM (Dittenbach 2000) which developed by G(H)SOM team of Vienna University of Technology to overcome the SOM's shortcomings where the map size has to be defined prior to training. G(H)SOM starts with the single-unit SOM at layer 0 with grid of 2x2 and it is adding the layers either a new row or a new column by referring to mean quantization error of a map.

Main storyline construction

The goal of main storyline construction is to analyze the topic structure and specify the main theme. At first, each event is weighted by Eq (1), which takes the ideas of genus and differentia words with previous *k* events modified from (Uramoto et al. 1998). It gives a high weight to the differentia words which previously do not appear because they contain new information. The edge (relationship) between a pair of vertices (events) is drawn when the similarity between events exceeds threshold.

Table 1. Summary of the pre-process

Preprocess	Description
Corpus collection	Collect a set of stories which belong to a certain topic by a news crawler robot
Word segmentation	Identify the word boundary in Chinese sentences by CKIP (http://ckipsvr.iis.sinica.edu.tw/)
Feature Filter	Filter the terms based on the criteria of *tfidf* and part of speech (POS)
Morphological Analysis	Unify terms which present the same meaning
Vector space export	Weight the terms by *tf* and location, and then covert them into vector space

$$weight\left(term_i^{\varepsilon_j}\right) = \frac{C_{\varepsilon_j}(term_i)}{\sum_{vector} C_{\varepsilon_j}(term)} \times \log \frac{k}{N_k(term_i)} \times g_{term_i}^{\varepsilon_j} \quad (1)$$

$$g_{term_i}^{\varepsilon_j} = \begin{cases} 1.5 & term_i \text{ does not appear in the previous } k \text{ events} \\ 1 & Otherwise \end{cases}$$

In Eq (1), $C_{\varepsilon_j}(term_i)$ denotes the frequency of $term_i$ in event ε_j, and $N_k(term_i)$ is the number of events that contain $term_i$ in previous *k* events. The constant value *k* limiting the link will exist with previous *k* events. The limitation is similar to the concept of *nearest parent* (Salton et al. 1988) that events are influenced by another event which occurs closely before them. *Cosine coefficient* (Salton et al. 1988) is used to calculate the similarity between events. The topic structure is constructed after no more links' similarity is below the threshold.

Given a topic structure, the next problem is how to find the main storyline. Intuitionally, the main topic will be discussed and distributed in events. Terms common to a topic will

repeatedly appear in most of stories. On the contrary, terms used to describe a unique event will appear only in certain events. Therefore, we define the main storyline as a path of events where the topic terms occur in high frequency. Topic terms similar to topic signatures (Lin et al. 2000) are a set of related words organized around head topics. Topic terms in this study will be determined by *document frequency*. Terms with high document frequency means that they are frequently discussed in most of stories and highly associated with the topic.

The algorithm for generating a maximum spanning tree (MST) to denote relevant events is taken to trace the main storyline. After obtaining topic terms, we use the relevance algorithm shown in Eq (2) to measure how relevant events are related to the topic set. It adopts the concept of *cluster-based retrieval* (Jardine 1971). In Eq (2), *TT* denotes the set of topic terms. *N* is the number of stories the event has, and *n* is the number of stories that contain the topic terms in event ε .

$$R(\varepsilon \mid TT) = \sum_{t \in TT} \left(\frac{tf_t}{\sqrt{\sum_{vector} (tf)^2}} \times \frac{n}{N} \right) \tag{2}$$

The MST is simply the tree spanning the nodes which in total has the maximum weight. It is usually solved by a greedy algorithm. It can be applied to find a path which goes through the high relevant events. The spanning tree is derived from the graph whose edges are weighted by Eq (3). To avoid the bias that a greedy algorithm only considers the current node and finds a local optimal, the relevance of next path that event α has will be measured. It forced MST consider the average relevance of next paths. Finally, branches are found from the storyline generated by the MST. Branches are the nodes that have lower relevance in the storyline path. These nodes appear because they can bring more relevant nodes in next path.

$$I(\varepsilon_i, \varepsilon_j) = \alpha \times R(\varepsilon_i \mid TT) + (1-\alpha) \times \frac{\sum R(nextpath\varepsilon_j \mid TT)}{k} \tag{3}$$

Storyline-based summarization

In the final stage of topic retrospection, the main purpose is to provide a concise description for each event and to compose a summary for a news topic. This study adopts Accumulated Weight Summary (AWS) (Goldstein et al. 2000) to compose a summary. At first, in each event on the main storyline, we extract the sentences from the first *p* paragraphs. The sentences will be segmented by punctuation marks, *i.e.*, period, semicolon, or exclamation point. The reason why we choose only the first *p* paragraphs is based on the heuristic of *inverted pyramid* (Brooks B.S. 1996) that denotes that the first few paragraphs contain the most important information. With this reason, the overview of an event in a document occurs at the preceding paragraphs, which are candidates for summarization. LabelSOM (Rauber 1999) and *tfidf* are then adopted as weight heuristic to give a high weight to sentences at the preceding paragraphs with these terms.

Consequently, this study accumulates distinct features for a concept by Maximal Marginal Relevance (MMR) (Goldstein et al. 2000). MMR computes penalty measures based on similarity factors to avoid selecting redundant sentences. An accumulated weight score is given to each candidate sentence by counting the occurrence of key terms. These candidate sentences will be ranked by their scores, and the first three sentences as candidate sentences are selected to compose the summary of an event. Moreover, each pair of sentences is

measured by their mutual similarity to avoid the redundant information.

The ordering of sentences in one event follows the reporting date and the location of paragraph in the original story. After preparing the summary of each event, this study applies two general strategies, *chronological ordering* and *majority ordering* (Brooks B.S. 1996), which are generally used for multi-document summarization to compose a summary complying with the main storyline. Finally, this study adds time period that an event occurred in front of each paragraph to compose the topic retrospective summary. A news reader can easily acquire the sketch of a topic from reading the summary.

In summary, the topic retrospection proposed in this paper uses GSOM for clustering documents into events. The storyline construction starts with determining events' weights Eq (1), and then using cosine coefficient to calculate the similarity between events to sketch the topic. Eq (2) is used to calculate the relevance between events and the storyline. The maximum spanning tree is derived from the graph whose edges are weighted by Eq (3). Finally, this study adopts Accumulated Weight Summary (AWS) to compose a storyline-based summary.

Demonstrating *SToRe* System for ETC News Topic Retrospection

The *SToRe* (Story-line based Topic Retrospection) system was implemented with the topic retrospection process elaborated in Section 2. The news topic regarding the deployment of an electronic toll collection (ETC) system in Taiwan's freeway systems is used to demonstrate the *SToRe* system. Far Eastern Electronic Toll Collection Co. (FETC) was commissioned by the Taiwan Area National Freeway Bureau to set up the ETC service according to the national transportation policy.

817 news reports containing ETC related news reports in Chinese, ranging from September 1998 to April 2006, were collected from http://www.udn.com.tw, a Chinese electronic newspaper to test *SToRe*. Parameters used by *SToRe* in this test are listed in Table 2. 213 terms are extracted from these news reports as well as a time stamp to form a 214 term vector. *SToRe* identifies 20 events from the event identification module using GSOM clustering. A news reader may recall occurred events while he or she is reading a news report in a topic. Therefore, a topic retrospection system should be able to take a news report, called anchor news article, given by a news reader to generate a storyline according to the anchor news article. Here we use a news article selected from the news corpus as the anchor news report to denote the topic that users are interested. In this stage, we choose the news article on February 2, 2006 as an anchor news report. This news describes the initiative of the ETC project and the traffic congestion issues on national freeways. According to the anchor we retrieve the relevant news events and build the storyline. Figure 1 illustrates the storyline based on the anchor news article, where rectangles denote the main part in the storyline and the circles denote the branch. Table 3 lists events shown in this storyline.

Figure1 denotes 10 events in storyline, and only last three events 665, 602 and 557 are directed related to the anchor news which refer to the kickoff of ETC system and the side effect on traffic congestion. In event 665, it describes that FECT wanted to practice the ETC system during the Chinese New Year. However the government used the chaos on OBU (On-Board Unit) price and hasty preparation as reasons to reject this proposal. Event 602 talked about the complaints from the public after practicing the ETC on two national freeways, including inefficient usage of ETC lane, the overpriced OBU and inconvenience of paying the ETC toll fee. Event 557 describes the reaction of legislators and officials, for

example, they wanted FECT to increase the usage rate of ETC lane, and otherwise FECT would receive the penalty. Legislators criticized that the government should take the responsibility of the ETC project.

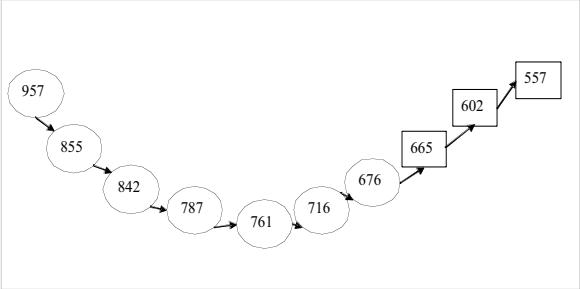

Figure 1. An storyline of launching ETC project

Table 2. The summary of parameters used for *SToRe*

Parameter	Value	Module	Criterion
Cycle times	1000	Event Identification	Empiric
Horizontal growth	0.008	Event Identification	Empiric
Learning rate	0.5	Event Identification	System initiation
k	3	Main storyline construction	Empiric
Similarity	0.0038	Main storyline construction	Median
Relevance	2.072	Main storyline construction	$\mu + 0.5 \times \sigma$
Similarity	0.6	Storyline-based summarization	$\mu + 3.5 \times \sigma$
p	2	Storyline-based summarization	Empiric

Table 3. A storyline of launching ETC project

Event #	Time duration	Event description
957	1998/09~2002/07	Discussion on the socio-technical environment of ETC system.
855	2003/10~2003/11	The formation of ETC consortiums
842	2003/12~2004/04	Discussion on selection of ETC consortiums
787	2004/09~2004/12	Discussion on lucrative ETC market and ETC investigation.
761	2005/01~2005/02	Companies cooperation with FETC
716	2005/08~2005/10	FETC tests ETC system.
676	2005/11~2006/01	Turmoil on OBU price debate
665	2005/12~2006/02	Discussion on the kickoff of ETC system
602	2006/02~2006/03	The complaints of ETC
557	2006/02~2006/03	Reflection from officials and legislators on ETC project.

Besides the events 665, 602 and 557, *SToRe* also extracts other events that may be relevant to or educe the main topic. The story started from the discussion of the ETC socio-technical environment, and then the government selected the bidding companies and the Far East

Group won the bid. Therefore, the Far East started to plan the ETC project. Many companies saw ETC as a lucrative market and wanted to cooperate with the Far East.

Storyline structure is very suitable for case study. Take Table 3 as an example, we can easily identify what happened to the ETC projects. We can read event 957 to figure out the original goals and expectation for ETC project. Then, from the formation of consortiums described in events 855 and 761, we know the cooperation strategy they adopted before and after the bidding. Moreover, from event 842 we can reexamine the ETC outsourcing process that caused criticism. Events 716, 676, 665,602 and 557 describe that the government and the FETC didn't coordinate very well, which resulted in many complaints from the public.

However, news readers may have different aspects of occurred events, and the proposed news topic retrospection possesses the flexibility for users to generate different summaries with different storylines according to the topic submitted by users. Figure 2 is another storyline generated by the *SToRe* based on the news regarding the scandal of ETC reported on April 4, 2006. Table 4 denotes the storyline with brief description. The storyline of ETC scandal has significantly different from the theme of launching the ETC project. Table 4 shows that the storyline focuses on the ETC bidding process which caused the investigation later. Moreover it shows that *SToRe* can reflect various aspects according to the user's preferences.

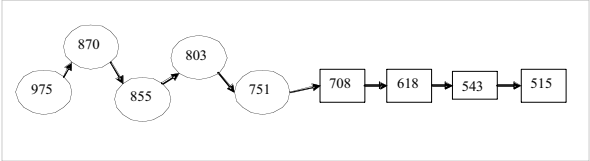

Figure 2. An storyline for the scandal of ETC project

Table 4. A storyline for the scandal of ETC project

Event #	Time duration	Event description
957	1998/09~2002/07	Discussion on the socio-technical environment of ETC system.
870	2003/07~2003/08	The announcement of inviting private companies to invest on ETC project
855	2003/10~2003/11	The formation of ETC consortiums
803	2004/06~2004/08	The suspecting ETC bidding process
751	2005/04~2005/07	ETC scandal on the conspiracy between government officers and business people
708	2005/10~2005/11	Discussion of prosecutors investigation
618	2006/01~2006/03	Discussion of public interest
543	2006/02~2006/03	The government's reflection about the ETC policy
515	2006/02~2006/04	Judicial verdict on FETC

Evaluation

In the evaluation, first we test if the *SToRe* is sensitive to the number of occurrence of 213 terms in anchor news articles. Second, we evaluate the overall precision and recall rates achieved by the *SToRe* with different anchor articles. Firstly, we compute the number of

terms out of 213 selected terms contained in each news article. In Figure 3, Y axis represents the number of articles, and the X axis denotes the average number of terms that an article contains. In this way, the number of articles that is close to $\mu-3\sigma$ or $\mu+3\sigma$ denotes that these articles contain much fewer or more selected terms, respectively. It also denotes that articles allocated closer to $\mu-3\sigma$ are less sensitive to selected terms, and vice versa. We randomly choose an article from each range as an anchor news article ranging from $\mu-3\sigma$ to $\mu+3\sigma$. Thus, in this experiment we obtain seven anchor articles: ETC_10.xml, ETC_69.xml, ETC_809.xml, ETC_327.xml, ETC_517.xml, ETC_581.xml and ETC_622.xml as candidate anchor articles for subjects to choose from.

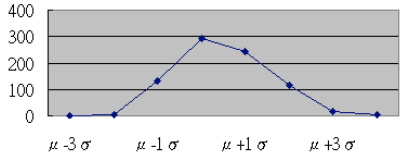

Figure 3. The average number of selected terms that articles contain

We then allocate fourteen subjects to select anchor news articles from corresponding number ranges of selected terms in order to evaluate the sensitivity of *SToRe* affected by the occurrence of topic terms. First, each subject selects an anchor news article from seven candidate news articles. Second, they read original news articles clustered in events and evaluate the relevance between the anchor news article and events. The relevant score ranges from 0 to 10. Each subject independently evaluates the relevance, and the average time that subjects spent to finish the evaluation is about 2 hours.

These seven candidate anchor news articles are also fed into the *SToRe* to generate seven corresponding storylines. Precision and recall are used to measure *SToRe*'s performance in generating individualized storylines for topic retrospection. Precision is defined as the ratio between the number of common events identified by the *SToRe* and human subjects and the number of events specified by the *SToRe*. Recall is defined as the ratio between the number of common events identified by the *SToRe* and human subjects and the number of events specified by human subjects. Table 5 summarizes the evaluation results on precision and recall.

Table 5. The precision and recall of the experiment

Anchor article	# of subjects	Precision	Recall
ETC_10.xml	2	6/9 = 0.667	6/11 = 0.545
ETC_69.xml	2	6/12 = 0.500	6/10 = 0.600
ETC_809.xml	2	6/10 = 0.600	6/10 = 0.600
ETC_327.xml	2	4/7 = 0.571	4/10 = 0.400
ETC_517.xml	2	7/10 = 0.70	7/8 = 0.875
ETC_581.xml	2	5/10 = 0.50	5/10 = 0.500
ETC_622.xml	2	7/9 = 0.778	7/8 = 0.875
Average	2	0.617	0.628
Standard deviation	0	0.104	0.182

Table 5 shows that the average values of precision and recall are 0.617 and 0.628, respectively. Since the standard deviations for precision and recall are relatively small (0.104 and 0.182 respectively), it indicates that the *SToRe* performance in terms of precision and recall is not significantly affected by inputting anchor articles containing various occurrences of topic terms. From this outcome, the *SToRe* can receive anchor news articles with various foci, and generate stably matched storylines. That is, the *SToRe* performance is robust and

insensitive to anchor articles with different occurrences of topic terms. In terms of recall rate, Table 5 shows that *SToRe* can retrieve more than 60% of events matching with individual readers' preferences, and it also identifies events that are directly related to the event indicated by the anchor article. Therefore, a storyline generated by the *SToRe* presents to a news reader more comprehensive view to understand the evolution of events according to his or her preference.

Discussion and Conclusion

This paper has designed and implemented a news topic retrospection system, called *SToRe*, to review occurred events under a news topic from a news corpus, and generate storyline-based summaries. *SToRe* consists of three major components: *event identification*, *main storyline construction* and *storyline-based summarization*. Techniques used for each component are novel or have been proved their fitness in text mining tasks. The study used Taiwan ETC (electronic toll collection) news topic retrospection to demonstrate and evaluate the *SToRe* performance. The results show that *SToRe* can generate robust storyline summaries by receiving anchor articles with various occurrences of topic terms. From the experiment, we also found that, in a generated storyline summary, 60% of retrieved events match with what specified by human subjects, and the rest of events describe causes or related incidents with the anchor article. Therefore, *SToRe* is an effective news topic retrospection system to generate individualized storyline summaries.

The contributions of this effort are mainly on the following three aspects: (1) it generates storyline structure to highlight major events which facilitate users to trace event evolution in a news topic; (2) it composes summaries from essential news sentences to mitigate general news readers' cognitive loads in digesting news reports; (3) it is robust to take various anchor news articles from users to generate individualized storyline-based summaries.

The capability granted by this study can extend further to many potential applications. For example, for business competitive intelligence, companies can collect news articles on occurred events related the company and its competitors to review their relations and interactions within or between industries under various aspects to obtain a concise summary. This may ease the cognitive loads while collecting information and making decisions. Moreover, *SToRe* can further be used for dynamic case studies as the supplementary materials for MBA education by automatically delivering business case retrospection to students to capture up-to-date case information. Our future research will further enhance *SToRe* system for these potential applications.

Acknowledgement

The authors would like to thank National Science Council, Taiwan to partially sponsor this study under the project No. NSC94-2416-H-007-007

References

Allan, J., Papka, R., and Lavrenko, V. "On-line new event detection and tracking," *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 37-45.

Berghel, H. "Cyberspace 2000: Dealing with information overload," *Communications of the ACM* (40), February 1997.

Brooks B.S., K., G.D., Moen, R., and Ranly, D. *News Reporting and Writing*, St. Martin's Press, NY, 1996.

Brown, J.S., & Duguid, P. *The social life of information*, Harvard Business School Press, Boston, 2000.

Dittenbach, M., Merkl, D. and Rauber, A. "The Growing Hierarchical Self-organizing Map," *Proceedings of IJCNN*, 2000.

Franz, M., and McCarley, J.S. "Unsupervised and supervised clustering for topic tracking," *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 310-317.

Goldstein, J., Mittal, V., Carbonell, J., and Callan, J. "Creating and evaluating multi-document sentence extract summaries," *Proceedings of the ninth international conference on Information and knowledge management*, 2000, pp. 165-172.

Ho, J., and Tang, R. "Towards an optimal resolution to information overload: an infomediary approach," *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*, 2001, pp. 91-96.

Jardine, N.a.V., R.C.J "The Use of Hierarchical Clustering in Information Retrieval," *Information Storage and Retrieval* (7), 1971, pp. 217-240.

Kohonen, T. *Self-organizing maps(2nd Ed.)*, Springer, New York, 1997.

Lin, C.Y., and Hovy, E. "The Automated Acquisition of Topic Signatures for Text Summarization," *Proceedings of the COLING Conference*, 2000.

McKeown, K.R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J.L., Sable, C., Schiffman, B., and Sigelman, S. "Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster," *Proceedings of the Human Language Technology Conference*, 2002.

Radev, D.R., Jing, H., Sty, M., and Tam, D. "Centroid-based summarization of multiple documents," *Information Processing and Management: an International Journal* (40:6), 2004, pp. 919-938.

Rauber, A. "LabelSOM: on the labeling of self-organizing maps," *Proceedings of International Joint Conference on Neural Networks*, 1999.

Salton, G., and Buckley, C. "Term-weighting approaches in automatic text retrieval," *Information Processing and Management: an International Journal* (24:5), 1988, pp. 513-523.

Sebastiani, F. "Machine learning in automated text categorization," *ACM Computing Surveys (CSUR)* (34:1), 2002, pp. 1-47.

Shih, J.Y., Chang, Y.J., Chen, W.H., Ho, J.H., and Kao, C.Y. "Constructing securities and futures markets legal maps of Taiwan using GHSOM," *Proceedings of 2nd International Conference on Digital Archive Technologies*, 2004, pp. 143-157.