Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2006 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2006

Topic Retrospection with Storyline-based Summarization on News Reports

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

Chia-hao Liang
National Sun Yat-sen University, chliang@iii.org.tw

Follow this and additional works at: http://aisel.aisnet.org/pacis2006

Recommended Citation
Lin, Fu-ren and Liang, Chia-hao, "Topic Retrospection with Storyline-based Summarization on News Reports" (2006). *PACIS 2006 Proceedings*. 75.
http://aisel.aisnet.org/pacis2006/75

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2006 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Topic Retrospection with Storyline-based Summarization on News Reports

Fu-ren Lin
Institute of Technology Management
National Tsing Hua University
Hsinchu City, Taiwan 300, R.O.C.
frlin@mx.nthu.edu.tw
Chia-hao Liang
Department of Information Management
National Sun Yat-sen University
Kaohsiung City, Taiwan 814, R.O.C.
chliang@iii.org.tw

Abstract

The electronics newspaper becomes a main source for online news readers. When facing the numerous stories of a series of events, news readers need some supports in order to review a topic in an efficient way. Besides identifying events and presenting the search results with news titles and keywords the TDT (Topic Detection and Tracking) is used to do, a summarized text to present event evolution is necessary for general news readers to review events under a news topic. This paper proposes a topic retrospection process and implements the SToRe system that identifies various events under a news topic, and composes a summary that news readers can get the sketch of event evolution in the topic. It consists of three main functions: event identification, main storyline construction and storyline-based summarization. The constructed main storyline can remove the irrelevant events and present a main theme. The summarization extracts the representative sentences and takes the main theme as the template to compose summary. The summarization not only provides enough information to comprehend the development of a topic, but also serves as an index to help readers to find more detailed information. A lab experiment is conducted to evaluate the SToRe system in the question-and-answer (Q&A) setting. The experimental results show that the SToRe system enables news readers to effectively and efficiently capture the evolution of a news topic.

Keywords: topic retrospection, Topic Detection and Tracking (TDT), event threading, summarization

1. Introduction

The prevalence of Internet technologies, such as World Wide Web (WWW), eases the information aggregation and dissemination. For example, online news spreads across Internet due to its responsiveness and customization features. In Taiwan, most of news databases provide the function of the clipping folder system (http://udndata.com/ndapp/specialtopic/StIndex) that collects related happened events in a topic by professional reporters. News readers can query news by accessing news databases, and edit individual clipboards. Even with those easy-to-query and personal clips, online news readers still face the information overloading problem. As the time passes by, people gradually forget events which occurred related to a specific topic, including what the most important event is, how it starts its turning point, and consequence.

Topic Detection and Tracking (TDT) sponsored by DARPA is a main task to manipulate and organize newswire stories. Topic detection is the task of grouping articles corresponding to the same topic, and topic tracking is the process of monitoring a stream of news stories to find those that track (or discuss) the same event. Although TDT can identify events in a topic, the results are presented as clustered news sets with titles and keywords but lack of an overview presentation to news readers.

In addition, researches of event threading beyond TDT consider the dependent

relationship among events (Nallapati et al. 2004). It views the structure of events and their dependencies in a news topic as an event threading problem. Although that event threading further analyzes the relationship, it dose not present a convenient way for news readers to easily understand the development process of news events. A news reader is more familiar to read news articles rather than a graph which presents the development of a news topic. Besides, most of researches (Shih et al. 2004; Smith 2002; Swan. and Allan 2000) use a set of keywords to highlight the themes of a clustered news set. Because of lacking the semantic level of characteristic, users need to look into individual documents to obtain the overview of events. Moreover, keywords present fragmental information, people need to compose meaningful sentences by their domain knowledge which is difficult for readers who are unfamiliar with the topic.

This study proposes a topic retrospection mechanism that demonstrates the story telling capability for news readers to understand the context of topic development in an efficient way. The implemented *SToRe* (Storyline-based Topic Retrospection) system consists of three main functions: *event identification*, *main storyline construction* and *storyline-based summarization*. It identifies various events under a news topic and constructs the relationship to compose a summary which gives readers the sketch of event evolution in a topic. The constructed main storyline can remove the irrelevant events and present a main theme that exhibits the major evolution of a news topic. Taking the main storyline as the template, the summarization not only provides enough information to comprehend the progress of a topic, but also serves as an index to help readers to find more detailed information.

A lab experiment is conducted to evaluate the *SToRe* system in the question-and-answer (Q&A) setting. The experimental results demonstrate that the *SToRe* system can help news readers effectively and efficiently capture the development of a topic, and in turn, reduce their information loads. Due to the length limitation of this paper, related research works on TDT, event trending, and text summarization will not be reviewed in a dedicated section; instead, they will be referred along with elaborating the topic retrospection mechanisms.

2. Topic Retrospection

2.1 Definitions

This study adopts the definition of TDT in topic and event (Allan .et al. 1998; Franz and McCarley 2001). A topic is defined as *a seminal event or activity, along with all directly related events and activities*, *e.g*., terrorism activity. An event is *something* (*non-trivial*) *that happens at a particular time and place*, *e.g*., Oklahoma City bombing and September 11 attacks. 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 et al. 2002; Radev 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 to give users the sketch of event evolution.*

In order to reduce the complexity of real world, 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) every story must belong to one of the events in ε, (2) each story can only belong 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.

2.2 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 1982), called GSOM (growing self-organizing map) (Dittenbach et al. 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.

2.2.1 Event identification

The first task in topic retrospection is to distinguish how many events happened in a topic. In previous researches (Dittenbach, et al. 2000; Shih et al. 2004), the self-organizing maps (SOM) (Kohonen 1982) 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 et al. 2000) to overcome the SOM's shortcomings where the map size has to be defined prior to training.

Table 1. Summary of the pre-process

Process	Description
Preprocess	
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

2.2.2 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 and Takeda 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.

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

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

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 and Buckley 1988) that events are influenced by another event which occurs closely before them. *Cosine coefficient* (Salton and Buckley 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 and Hovy 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 main storyline. It adopts the concept of *cluster-based retrieval* (Jardine et al. 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 number of next path that event ε_j has will be measured. It decreases the probability that the MST finds a short path that with few events but high relevant covered.

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

Finally, branches are added to the main path generated by the MST. This study only considers nodes which connect to the main path and their relevance exceeds a given threshold. It enhances the storyline construction by remedying what the MST can only build a path without branches.

2.2.3 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 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 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 et al. 2002), 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 *SToRe* system uses GHSOM 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.

3. *SToRe* System Evaluation

The *SToRe* (Story-line based Topic Retrospection) system is implemented with the topic retrospection process elaborated in Section 2, and inputs a news corpus to obtain preliminary results. Lab experiments are conducted to evaluate its performance in a question-and-answer (Q&A) setting.

3.1 Preliminary results

In order to properly evaluate *SToRe*, this study selects a news corpus based on the following criteria: (1) the number of stories exceeds readers' cognitive loading, (2) the time period that the news topic occurred is far away enough that readers may not remember anymore, (3) there is no similar topics occurred that readers can reference, and (4) readers are not familiar with the topic or lack of domain expertise.

Experimental data sources are collected from the business and economics categories of *Udndata.com* clipping folder system. Udndata.com (http://udndata.com) which established by United Daily News Group (UDN) is a Chinese news retrospective database. These categories are the most dissimilar to subjects. There are 188 topics in these two categories and the average number of stories is 69. According to the aforementioned selection criteria, "*HP's acquisition of Compaq*", is picked as the news topic. It occurred during 2001 and 2002 (news reports from September 1, 2001 to Arpil 1, 2002) and is far away from the time, June, 2005, that we conducted the experiment. Moreover, subjects cannot read all 179 stories in short time. Therefore, the selected topic conforms to these criteria.

Figure 1 shows the main storyline of "*HP's acquisition of Compaq,*" where the solid lines are constructed by MST and dashed lines are branches. The brief introduction in block is labeled by human to explain the sketch of event evolution the *SToRe* system produced.

3.2 Experimental design

The experiment is conducted with two groups under nearly identical conditions, and each group will be presented by different formats of topic retrospection. Subjects include 49 graduate students from a Department of Information Management of a Taiwan's university. Among them, 73% and 27% students are male and female students, respectively. Every subject is randomly assigned into the experimental or control group according to the sequence s/he registers to the experiment. Thus, subjects in both groups major in the same field with similar educational background. All subjects receive NT$ 200 to compensate their time and efforts to participate to the experiment. In order to encourage subjects to make the best efforts on answering questions, additional prizes were rewarded to subjects who were ranked in the five highest scores.

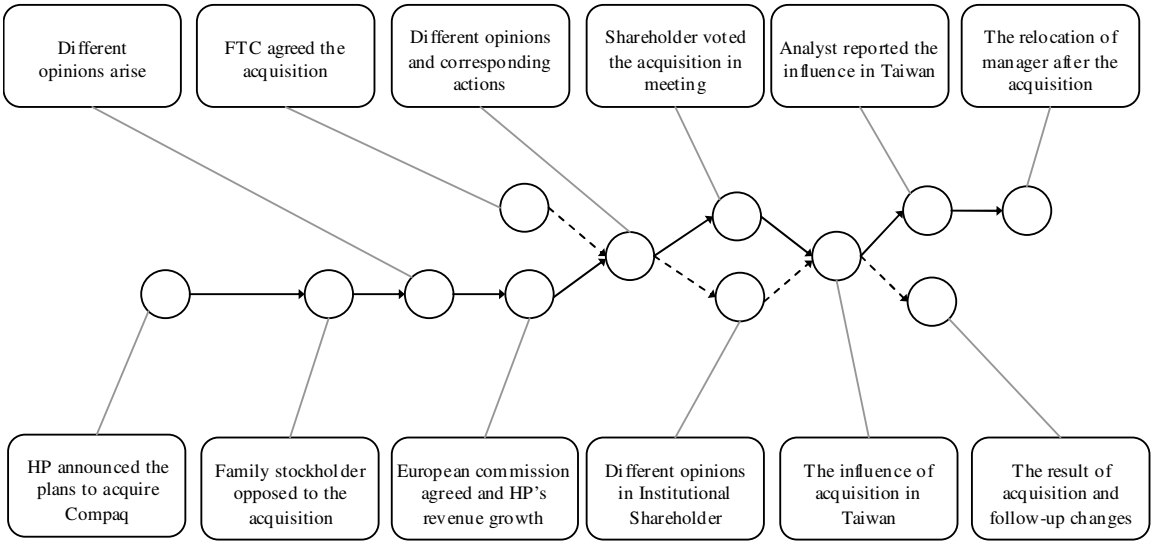

Figure 1. Event revolution process of "HP's acquisition of Compaq"

The experimental procedure of lab experiments in a Q&A setting is summarized in Figure 2 (Mani and Maybury 1999). It takes 40 minutes for a subject to finish the experiment. The main differences between the experimental and control groups are on the display

format of topic retrospection. The experimental group will read a storyline-based summarization, and the control group will see the result of events identified with news title lists. Thus, the experiment compares the proposed *SToRe* system and the traditional TDT on helping news readers understand the evolution of a news topic.

As illustrated in Figure 2, the experiment consists of three phases. Before starting the experiment, subjects are assigned to read documents to get familiar with the system. In order to eliminate the bias from the different degrees of familiarity with the system, in Phase one, subjects will practice the system with a simulated examination, Chen-Soong Meeting, the topic related to the historical meeting between President Chen and one opposite party chairman Dr. Soong in Taiwan. Each subject is allowed to read information in 2 minutes and answer question in 3 minutes. Phase two is the formal examination that each subject is asked to read the text provided by the system in different formats for 9 minutes, and questions about the topic will be prompted to test subjects' degree of understanding for 20 minutes. In Phase 3, each subject will fill out a questionnaire which collects subject profiles in reading electronic newspaper and their perception to the *SToRe* system.

3.2 Experimental questions

Whereas the design of examinational questions is a critical part in the experiment, a domain expert who does not involve with this study authorized questions. The domain expert was given the full news text of the topic without reading the storyline-based summary. According to the characteristic of news topic, questions are categorized into five types: *WHO*, *WHAT*, *WHEN*, *WHICH* and *OTHER*. The scope of questions includes different levels from narrow in one event to broadly cover several events. Furthermore, some questions are considered the relationship between a pair of events or the sequence to sort the events. The format of questions include multiple-choice, blank filling, and true-false items.

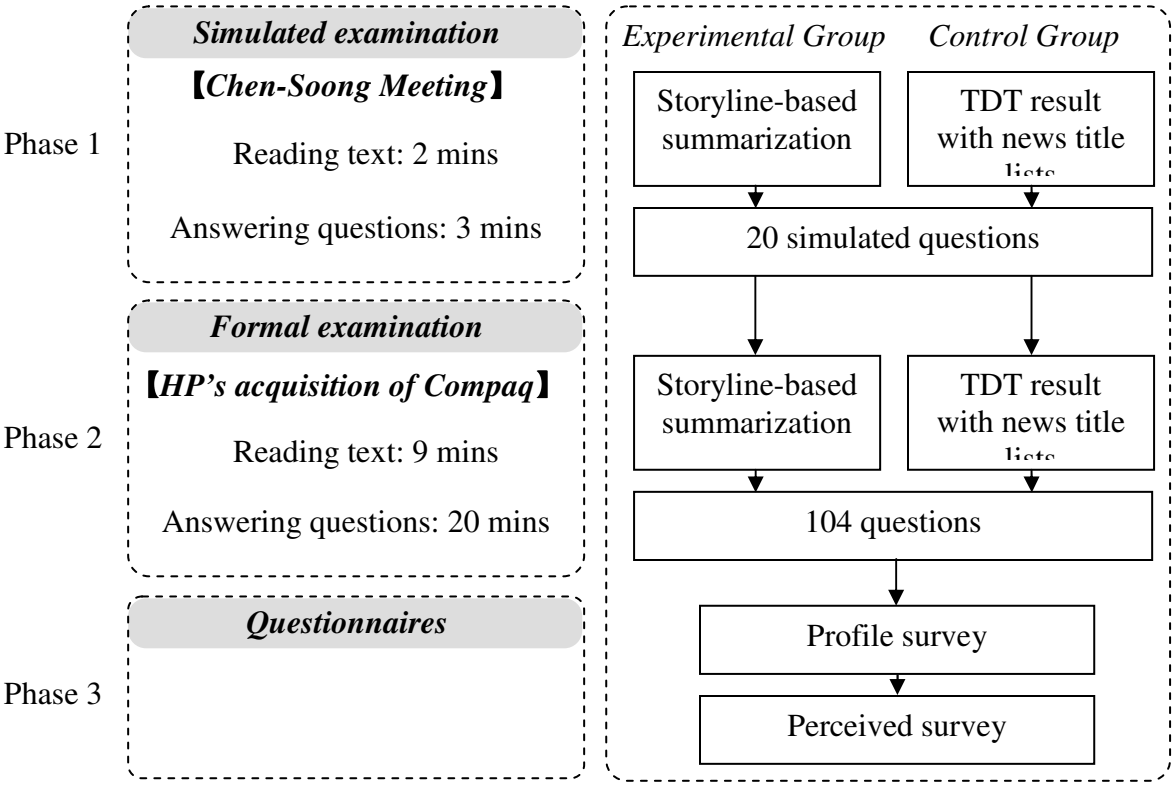

Figure 2. Experimental procedure

Table 1 shows the distribution of 104 questions in different types and Table 2 shows their distribution in different question formats. 26 questions are set as a unit, and each unit is composed of different types of questions proportionally to the percentage of types in these 104 questions. The main purpose of this question dispatch is to assure that every type of questions has the chance to be answered in one unit. But the questions in the category of *OTHER* will only be assigned into the last unit because they are less relevant to the topic and hard to classify. The ordering of questions in each unit will be sorted randomly. Hence, the sequence of questions a subject faces is different from each other. It eliminates the bias that only easy or hard questions are answered by subjects firstly. After the experiment, the study adopts human-scored method by two parallel reviewers to verify the accuracy.

Table 1. Types of questions in formal examination

	WHO	WHAT	WHEN	WHICH	OTHER	Total
Questions	12	41	13	25	13	104
Percentage	11.5%	39.4%	12.5%	24.1%	13.5%	100%

Table 2. Format of questions in formal examination

	Multi-choice	Blank filling	True/Fase	Total
Questions	46	42	16	104
Percentage	44.2%	40.4%	15.4%	100%

4. Experimental Results and Discussion

Results from 48 subjects are valid. One subject is removed from the result set because the subject failed to follow the experimental instruction throughout the process. We use the confidence interval $\alpha = 0.05$ to test the statistical significance.

At first, the time, frequency, and category that subject read electronic newspaper are asked in the experiment to understand subject profiles. From the statistical results of reading frequency (p-value = 0.785), reading category (p-value = 0.892), reading time (p-value = 0.159) and gender (uniform distribution), the subjects profiles are not insignificantly different. Thus, this experiment was conducted in similar subject profiles.

Moreover, the hypothesis of guessing strategy is verified to assure that subjects made the best efforts during the experiment. If subjects adopt the guessing strategy, they will guess the answer without finding any information from text. The action cannot reflect the real impact of different display formats. Hence, a multi-choice or true-false question is regarded as a guessable question. On the other hand, a blank filling question is categorized as a non-guessable question. The results are not significantly different (p-value = 0.127); therefore, there is not enough evidence to support that subjects use the guessing strategy for answering questions during the experiment.

In terms of effectiveness, we expected that *SToRe* could help news readers correctly capture the theme of news articles of a topic in short time. The score, correctness rate, and answered questions in a Q&A test are used to evaluate the *SToRe* effectiveness. Table 3 shows that subjects in the experimental group significantly outperform those in the control group in terms of correctly answering questions indicated by the p-value of score (0.012) and correctness rate (0.028). This outcome supports that *SToRe* effectively brings news readers a sketch of topic in short time. Although the number of answered questions for subjects in the experimental group is larger than that in the control group, there is no significant improvement. This outcome can be explained as follows. *SToRe* gives subjects an overview that some events occurred. When a subject faces a question, he/she may verify his/her perception to ensure the correctness of his/her answer by reading the indexed news articles. Therefore, two groups spent the similar length of time in one question.

Table 3. Statistical results in terms of effectiveness

Measurement	Control group	Experimental group	p-value
Score	9.29	14.50	0.012*
Answered questions	18.25	23.46	0.129
Correctness rate	0.52	0.63	0.028*

In terms of efficiency, the number of clicks to articles prior to answering a question is recorded in order to understand the efforts a subject spends on answering questions. If *SToRe* provides an efficient channel to find information, a news reader may find answers in few steps. The summarization is treated as an index to help a news reader to construct his/her own knowledge structure regarding a news topic. Table 4 shows the result that subjects in the experimental group used significantly less clicks than those in the control group ($p = 0.012$).

In addition, this study analyzes the effects of different hyperlinks to news articles in the experimental group. In the experimental settings, two ways that subjects can link to original new articles are designed. One is called LFS (Link-From-Summarization) that the news article is linked from the end of each sentence in the summary to the corresponding article. The other is called LFL (Link-From-List) that news articles are linked from a list of news titles corresponding to a paragraph of the summarization. We hypothesize that subjects will prefer LFS to LFL if the summarization plays the indexing role. A paired samples *t*-test was adopted to examine whether this hypothesis was supported. In Table 4, *p*-value (0.023) is less than confident level, and the mean of LFS (23.87) is greater than that of LFL (9.42). It indicates that links from summarized sentences facilitate the search of answers.

Table 4. Statistical results in terms of efficiency

Measurement	**Control group**	**Experimental group**	***p*-value**
Articles clicks	49.67	33.29	0.012*
Measurement	**LFS**	**LFL**	***p*-value**
Hyperlink clicks	23.87	9.42	0.023*

Finally, this study uses 5 point Likert scale to measure subjects' perceived values of *SToRe*, and the outcomes are shown in Table 5. The means are all above the average, 3.0. The highest mean is usefulness, and the lowest mean is ease of use. From these results, we found that *SToRe* is very useful for subjects who regarded it as an efficient channel. It is the same result with our statistical hypothesis in Table 4 (article clicks and LFS). In terms of ease of use, this study is the first time that subjects operate such a system as *SToRe* to capture the evolution of events in a news topic. Hence, subjects couldn't comprehend functions embedded in the system. By conducting training activities, we believe that subjects may raise the perception of ease to use.

Table 5. Statistics of survey in the experimental group

Measurement	# of subjects	**Mean**	**Std. deviation**	**Std. error mean**
Usefulness	24	3.708	1.083	0.221
Ease of use	24	3.125	1.191	0.243
Efficiency	24	3.500	1.142	0.233
Satisfaction	24	3.250	0.944	0.193
Comprehensibility	24	3.250	0.897	0.183

5. Conclusion and Potential Applications

This study proposes a mechanism to help a news reader to review a news topic in short time. Comparing with the previous approaches that only identify events and lists with news titles and keywords, the *SToRe* system adopts an integrated framework of TDT, event threading and summarization to support topic retrospection. It considers the quantity, format and quality of information to mitigate the information overloading problem faced by news readers.

The *SToRe* system can be applied in many fields. The most direct application is in news

database. In Taiwan, most of news databases provide the function of the clipping folder system that collects related happened events in a topic. Hence, the mechanism may help a news reader to get a sketch of happened events in a topic, and facilitate the reader to select specific news articles to read. It may also be applied to time-stamp documents for companies or government agents. They can reserve digital documents of executed projects and capture the project development process by using the proposed topic retrospection mechanisms. The summary of a project can be used as a guide for a freshman or a junior manager to plan similar projects in the future. They can learn from these occurred activities in similar projects and pay attention on critical activities which lead a project to succeed or fail.

The proposed methodology can be also used as a content analysis supporting tool for researchers who are used to adopt the content analysis methodology and attempt to analyze a large set of documents. It is quite laborious and time consuming to manually analyze a large document set. However, by using techniques such as topic retrospection proposed in this study, researchers can get a better focus and index to access documents to conduct further analysis. It is beneficial for researchers who are dealing with text generated via Internet, such as Weblogs and newsgroups.

References

Allan, J., Papka, R. and Lavrenko, V. "On-line New Event Detection and Tracking," *Proceedings of the 21sl ACM SIGIR,* 1998.

Barzilay, R., Elhada, No. and Mckeown K.R. "Inferring Strategies for Sentence Ordering in Multidocument News Summarization," *Journal of Artificial Intelligence Research,* (17), 2002, pp.35-55.

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

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," In *Topic Detection and Tracking Workshop,* 2001.

Goldstein, J., Mittal, V., Carbonell, J. and Callan, J. "Creating and Evaluating Multi-document Sentence Extract Summaries," *Proceeding of CIKM'00,* McLean, VA, 2000.

Jardine, N. and Van, 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-Verlag Berlin Heidelberg New York, 1982.

Lin, C.-Y. and Hovy, E. "The Automated Acquisition of Topic Signatures for Text Summarization,". *Proceedings of the 17th conference on Computational linguistic,* 2000.

Mani, I. and Maybury, M. "Introduction,". In *Advances in Automated Text Summarization*, I. Mani and M. Maybury (eds), MIT Press, 1999, pages x—xv,

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*, San Diego, CA, 2002.

Nallapati, R., Feng, A., Peng, F. and Allan J. "Event Threading within News Topics," *Proceedings of CIKM,* 2004.

Radev, D.R., Jing, H., Stys, M. and Tam, D. "Centroid-based Summarization of Multiple Documents," *Information Processing and Management*, (40), 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., Buckley, C. "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing & Management,* (24:5), 1988, pp.513-523.

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," *International conference on digital archive technologies*, 2004.

Smith, D.A. "Detecting and Browsing Events in Unstructured Text," *Proceedings of the 25th Annual ACM SIGIR Conference*, 2002, pp.73-80.

Swan, R. and Allan, J. "Automatic Generation of Overview Timelines," Technical Report IR-198, University of Massachusetts, Department of Computer Science (CIIR), 2000.

Uramoto, N. and Takeda, K. "A Method for Relating Multiple Newspaper Articles by Using Graph, and its Application to Webcasting," *Proceedings of 36th conference on Association for computational linguistics, 1998.*