Time-based Contextualized-News Browser (T-CNB)

Akiyo Nadamoto
National Institute of Information and Communications Technology
Hikaridai, Seikachyo, Kyoto, Japan
nadamoto@nict.go.jp

Katsumi Tanaka
Kyoto University
Department of Social Informatics, Graduate School of Informatics
Yoshida Honmachi, Sakyo-ku, Kyoto, Japan
ktanaka@i.kyoto-u.ac.jp

ABSTRACT

We propose a new way of browsing contextualized-news articles. Our prototype browser system is called a *Time-based Contextualized-News Browser* (T-CNB). The T-CNB concurrently and automatically presents a series of related pages for one news source while browsing the user-specified page. It extracts the past related pages from a user-specified news articles on the web. The related pages outline the progress of user-specified news articles. We call the related pages 'contextual pages'. Using the T-CNB, a user only needs to specify one news article on the web. The user then automatically receives past related news articles, which provide a wider understanding of the topic. The T-CNB automatically generates and presents contextualized news articles.

Categories and Subject Descriptors

H.5.2 [**User Interface**]: Windowing Systems; I.7.m [**Document and Text Processing**]: Miscellaneous

General Terms

Design, Documentation

Keywords

contextualized news articles, web browser, topic graph

1. INTRODUCTION

Currently, a huge volume of world news is reported on the web. News stories can be reported anytime, and can also change anytime. However, if you miss reading the news for a few days, it is difficult to grasp the progress of current stories. If we want to understand the progress of today's news stories, we have to search for related previous articles, and open and read them. We may also have to repeat this operation and we may have to read multiple time-series news articles to understand the past progress of today's news. Some news sites present related news article anchors on their news pages, but we have to open and read them, and also have to repeat the operation. We cannot get related news articles simultaneously, and the process is tedious. We consider that it would be more convenient if a browser could concurrently and automatically present the user with news stories and their progress. In this paper, we describe a new browser called a *Time-based Contextualized-News Browser* (T-CNB). The T-CNB extracts related web pages from user-specified web pages and concurrently and automatically presents a series of past pages relating to one news source. The related pages outline the progress of user-specified news articles. We call the related pages 'contextual pages'. There have been many studies on news articles,[1],[2],[3]. The objective of much of them are to summarize multiple documents, but our research focuses on the computing of structure-based topics and topic graphs.

WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

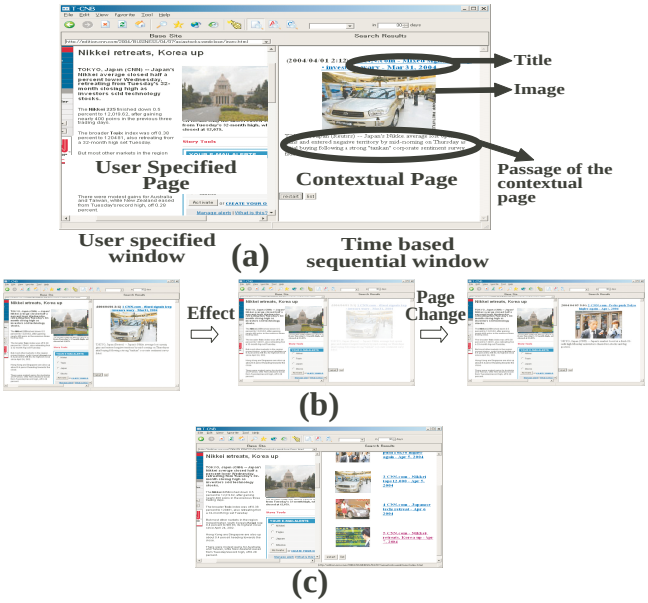

Figure 1: User Interface

The outline of the T-CNB is as follows; first, users specify news articles that they want to read on the news site. Then, the T-CNB searches for contextual pages on the same web site based on the topic graph for the user-specified page or pages that have been searched. In this time, the T-CNB finds contextual pages that have different passages from similar web pages by using a topic graph. These pages then become candidates for contextual pages. Next, the system presents the passage of the contextual page automatically from oldest to latest by using passive viewing method.

2. USER INTERFACE

Figure 1(a) shows an image of the T-CNB user interface. The left window shows a user-specified page and the right window shows a component of a contextual page. The T-CNB automatically presents contextual pages one by one, which users can without any interactions. A news source consists of a series of news articles on a time axis. Each contextual page presents a point in the whole of the

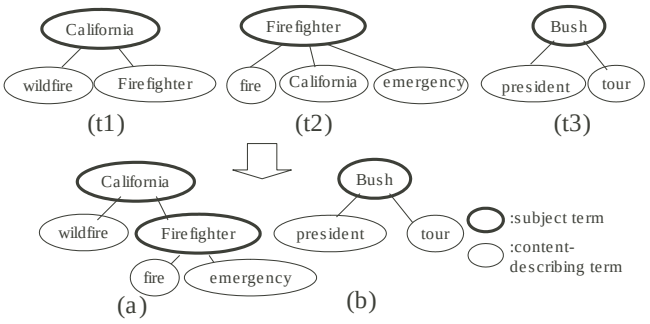

Figure 2: Example of Topic Graph

news source. Users can understand the outline of a news source by connecting each contextual page in the series. In our system, contextual pages are presented automatically in time order in an animated Microsoft PowerPoint display. There is a fade and zoom effect when a page changes (Figure 1(b)). The T-CNB presents a list of the contextual pages when the display ends. This list consists of images and titles. When a user clicks the repeat button, the system restarts the animation. Clicking on a title brings up the web page for the contextual page (Figure 1 (c)).

3. CONTEXTUAL PAGE

It is necessary for the T-CNB to present similar pages that have meaningful differences. A news article has a main topic and several sub-topics. Between related pages, the main topic may become a sub-topic or a sub-topic may become the main topic. We compare the change value of a set of topics by using a topic graph. Our concept of context relates to the similarities and differences between topic sets. In this way, a contextual page contains both similar and different topic sets. The T-CNB extracts contextual pages using a topic graph. The topic graph is based on the topic structure, which consists of subject and content terms.

3.1 Extracting Topic Structure

For given page P, its topic t_i, $i \in \{1, \ldots, n\}$ is simply represented as a pair of *a subject term* s_i and a set C_i of *content-describing terms*. C_i consists of multiple *content-describing terms* c_{im}, $m \in \{1, \ldots, k\}$. s_i is noun and its term frequency is more than the threshold α. c_{im} in a given page is intuitively the term that has a high cooccurrence relationship with s_i in the page. A web page P may have more than one topic, and so, s_i is associated with multiple c_{im}.

3.2 Extracting Query Keywords

Generating topic graphs

The T-CNB generates a graph from a web page by using the co-occurrence relationship among extracted subject terms and content-describing terms. We call the graph a topic graph. The topic graph is a undirected graph, in which each node corresponds to an extracted subject term or a content-describing term. There are usually multiple connected components in the topic graph. Figure 2 shows the T-CNB topic graph. In this case, the subject terms "California" and "Firefighter" co-occur. The graph joins them, but the subject term "Bush" does not co-occur with any other subject terms. Then, the topic graph for page P consists of two connected component graphs. Thus, the system creates a topic structure for the web page.

Extracting query keywords

In the topic graph, the connected component that includes the largest number of terms (vertices) becomes the main topic of a web page; other connected components become sub-topics. The T-CNB extracts query keywords from the main topic. The subject terms and content-describing terms of the main topic become query keywords. Thus, the query keywords Q for the current page P becomes $Q = (s_1 \wedge s_2 \wedge \cdots \wedge s_n) \wedge (c_1 \vee c_2 \vee \cdots \vee c_m)$. Here, $s_1, \cdots, s_n$ and $c_1, \cdots, c_m$ are the subject terms and content-describing terms contained in the maximum connected component.

3.3 Selecting contextual pages

Our hypothesis for selecting contextual pages for page P is (1)The main topic of page P should have a high similarity to both of the main topic and sub-topics of the contextual page (2)The sub-topics of page P should have a low similarity to both of the main topic and sub-topics of the contextual page. The system extracts web pages that have similar main topic and some variation in sub-topics.

Similarity-detection

The T-CNB computes the degree of similarity between the main topic of a current page and the connected components of past pages. Let G and G' be topic graphs for a current page P and a past page P', respectively. The similarity degree of G' with regard to G is defined as follows:

$$Sim(G', G) = \frac{1}{m} \sum_{j=1}^{m} \frac{\mid v(G'_j) \cap v(G_1) \mid}{\mid v(G'_j) \cup v(G_1) \mid}$$

where G_1 and G'_1 correspond to the main topics (connected components having the largest number of vertices) of G and G'respectively, G_i and G'_j denote other topics, and $v(G)$ denotes the set of all the vertices of a graph G.

The T-CNB computes the above similarity degrees of retrieved past pages with regard to a current page P, and selects only the past pages as contextual pages for P whose similarity degrees are greater than a given threshold γ.

Difference-detection

The difference degree of G' with regard to G is defined as follows:

$$Diff(G', G) = 1 - \frac{1}{(n-1)m} \sum_{i=2}^{n} \sum_{j=1}^{m} \frac{\mid v(G'_j) \cap v(G_i) \mid}{\mid v(G'_j) \cup v(G_i) \mid}$$

If $Diff$ is greater than the threshold δ, the page becomes a candidate for a contextual page.

4. CONCLUSION

We have described our Time-Based Contextual Browser (T-CNB). The T-CNB presents a user-specified news page and related pages concurrently and automatically. It extracts related pages to the user-specified web page using topic graphs, which consist of the topic structure, and subject and content-describing terms. The T-CNB compares connected components in the topic graph, finds similarities and differences between each component, and extracts contextual pages. Moreover, the T-CNB presents concurrently and automatically the related pages a series of one news source when user read he/she specified page.

5. REFERENCES

[1] M.Spitters and W.Kraaij, "A Language Modeling Approach to Tracking News Events," TDT 2002 Evaluation workshop, Gaithersburg, MD, USA, 2002.

[2] J.Allan, R.Papka, V.Lavrenko,"On-line New Event Detection and Tracking", Proc. of 21st International ACM SIGIR, pp.37-45, August 1998.

[3] Columbia's NewsBlaster site homepage http://www1.cs.columbia.edu/nlp/newsblaster/