2004－ICS－138　(16)
2004／12／5

Document Retrieval through Time-dependent Events in the Web

MIKIHIKO MORI[†] and SEIJI YAMADA[††]

We propose document retrieval in the Web by a visual interface arranging stories about events according to time and user's context. We present a concept of representation time-dependent events in the Web. We give the definitions of the terms of story, event and topic. It is necessary to illustrate Web documents as timeline to retrieve stories through a time. We also present the chronological visual interface to interact with users. The interface shows stories in a timeline, which is made of Web documents. The visualization in the interface shows clusters of stories that are similar mutually.

1. Introduction

There are a large number of pages in the Web for reasons that most people are thinking that the Web is a place of information service and communication. A kind of Web pages tells news in the real world as online news sites, another teaches know-how or knowledge of something. In recent years, it is increasing that people begin to communicate through the Web so that Web pages become more functional. They publish diaries or weblogs (blogs) on the Web for talking about that they thought, felt and found something.

Many researchers challenged to solve finding necessary information in such pages. As a solution, many search engines exist for retrieving the information. However, these search engines tend to present a lot of hit pages by users' queries. As one technique of more precise page ranking, Brin and Page proposed link-based estimation of ranking[2] and presented Google which is now a largest search engine. When a user wants to search through blogs, Nanno et al. proposed a search method specialized in blogs[5].

Although Web users can obtain precise information, a series of pages to which is related mutually can be hardly found out. For example, when a user is interested in the affair of the avian influenza told on the Web in 2003, the user may use a search engine by way of input a query like “bird flu 2003”, but the user obtains much information by a lot of pages from online news, blogs, diaries, public comments, etc. These pages is not clustered and not arranged in order of timeline. Though some search engines show the time at which the page was update, it is not satisfied. Some news sites present link pages for collecting and tidying up on some events. However, they provides just their inside pages.

We aim to develop a technique which retrieves several series of relevant events in the Web pages, clusters by searcher's viewpoint and visualizes the events and the series, and then to build a system implemented in the technique. In this paper, we propose our framework of technique.

2. Related Works

Some search engines attend to the last update time of the Web page or cluster similar pages of search results as topics. Fresheye[☆1] shows the last update times when it could get the times. Google[☆2] also shows the last update times if it assumes the pages are frequent updated. Although these search engines attend to document's times, they do not attend to timeline and topics. On the other hand, Clusty[☆3] used the clustering engine of Vivisimo[☆4] organizes search result pages. Each of the clusters labeled short phrase. WiseNut[☆5] is also a clustering search engine which has narrow down the results by the words in the label as additional feature. These search engines could enhance the results by topics but do not look at document's times.

Allan et al. proposed Topic Detection and Tracking (TDT)[1]. TDT aims at developing algorithm for detecting and tracking topic from newswire corpus which had already structured in SGML format. Although it is similar to our work, we attend to time-values in documents and do not use prepared corpus.

[†] Kyoto University
[††] National Institute of Informatics

[☆1] http://www.fresheye.com/
[☆2] http://www.google.com/
[☆3] http://www.clusty.com/
[☆4] http://www.vivisimo.com/
[☆5] http://www.wisenut.com/

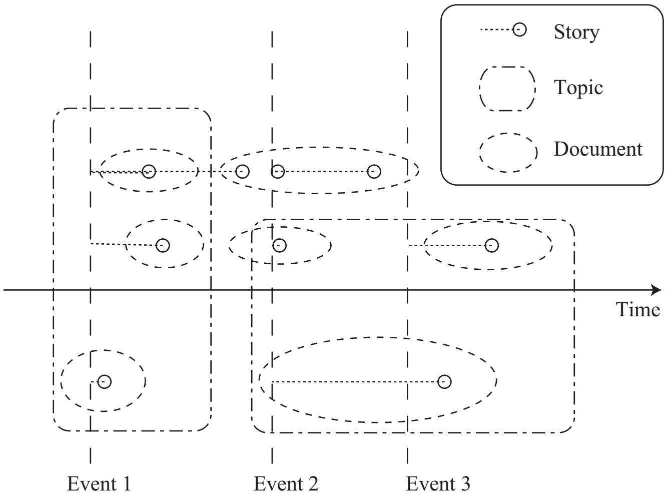

Fig. 1 The relation between events, stories and topics

There are works of time-dependent visualization techniques for personal information. Lifestreams[4] manages a time-ordered stream of documents that are stored in daily life of each user. The user is able to filter from the stream and to create documents to it. Time-Machine Computing[7] proposes a time-centric approach that information is organized and visualized user's documents. It enables various visualization of time in addition to storing and finding out their documents. Stuff I've Seen (SIS)[8] searches personal contents by the value of annotating timeline with event landmark like digital photographs. It provides timeline-based visualization of search results.

Kumar et al. uses visualizing timelines of metadata in digital libraries. Although they attend to metadata in closed libraries, we step in contents on the Web in which people maintain each other's pages[9].

Blog Watcher[5] searches blogs in terms of date and words as a query. It also shows burst of a word, which have a relation to a popular topic. It does not, however, show related documents by timeline while blog writers surge in a topic.

Dying Link system[6] displays freshness of Web pages by the appearance of a link. If a page was not modified more old, the link to the page is effected the look.

Loom[3] visualizes discussions of Usenet groups on a 2D map to show the social patterns. Each participant represent on an axis and posts of a participant are plotted dots along other axis as the time.

3. Time-dependent Event and Stories on Topics

We use the terms such as event, story and topic which are similar to the usage of TDT. Their definitions are as follows.

We consider a document to tell some stories. The story says an event occurred or regarded to occur in the world. The story sometimes describes some events. Each story must have its written, first appeared or update time shown in the document. The story also includes descriptions about the event and related events. A topic is in the context of a reader of documents, which usually contains the stories of a certain event. The topic frequently contains the stories of similar events or related events.

A time-dependent story among others has the times created or updated, or the risen time of an event. In this paper, we employ just the story depended on the time and omit the others.

Fig. 1 shows an example of relationship of stories, events and topics. All stories are derived from their individual events (shown as start point of dashed line to each story-dot). Original documents including stories represent dashed ovals. The left dashed rectangle describes a topic about single events. On the other hand, the right dashed rectangle describes another topic about two events which may be related or similar events in a context. The context is determined by each documents' reader.

4. Searching Time-Dependent Stories

Readers of Web documents often have an inter-

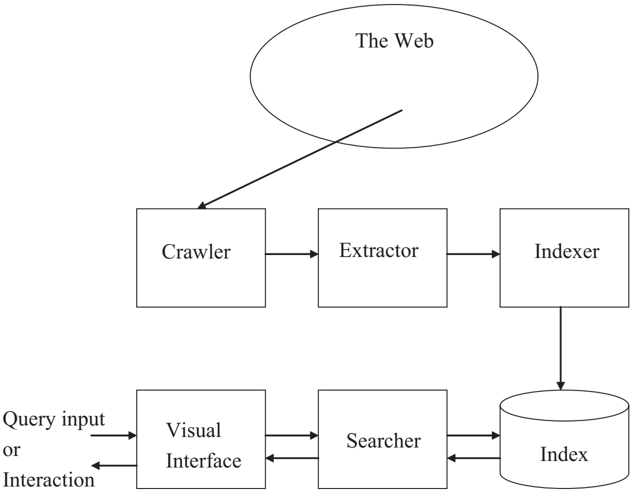

Fig. 2 The architectural concept

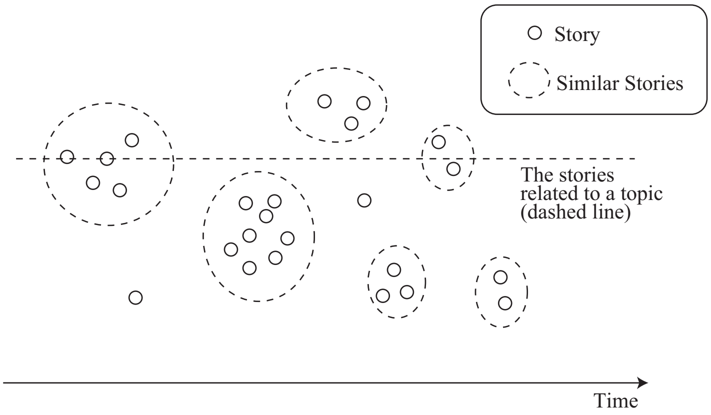

Fig. 3 Visualizing timeline in the interface to search and to explore

est in the time of the documents. To know a topic that they meet first, they need to see its context. They must usually search for stories on the topic. However, the following presentations are usually not shown for the readers:

- Freshness or recentness of the stories in a document describing the topic,
- Related topics which do not include the events of the original topic,
- The stories related to the topic in a document by extracting each stories which describe different topics,
- Stories in chronological order only on the topic,
- Relations among a lot of stories from different documents,
- To use reader's context to search the stories.

Our Approach to present timeline helps these demands of using documents on the Web. When search for stories on a topic, the timeline presentation shows the stories about the topic or stories of topics related to it. We propose the following techniques for time-dependent search and visualization.

Crawling is the first action to utilize Web documents. Crawlers get Web documents to follow each hyperlink in them. For time-dependency, our crawler will be led to the pages that have expressions of the time like online news, blogs and diaries. Here, a Web document is considered a Web page as a file.

Extracting of the time is necessary to present timeline of stories. Every story is made related to the times created or updated, or the risen time of event in the story. Every story is also extracted at same time.

Indexing makes from stories Web pages crawled. A story can properly be hit with words in the index. The searcher has to be able to distinguish each of the time exampled above.

Search uses above index by the times, words and other stories or topics.

Visual interface illustrates stories of search results and arranges stories by relations between stories, events and topics. Users interact through the interface to search and explore the stories on their topics.

Our concept of the architecture integrates these techniques as modules in **Fig. 2**. These modules cooperate to present the stories on a topic. The crawler gets Web pages via HTTP, retrieves hyperlinks and gets other pages over again. It also sends the pages to the extractor of the times and words. If the extractor recognizes the page including time-dependent stories, the extractor sends a message to the crawler back. When the crawler gets the message, it reinforces crawling around the page for the reason that the crawler is probably within online news, blogs or diaries. If the crawler gets a time-independent story, the extractor will estimate the time of its content. The indexer gets the times, the words, and the content of story from the extractor and adds them to the index optimally.

When users demand to view the stories on a topic, they input query about topic at first, then the interface shows the stories visually. These stories have been filtered by searcher beforehand. The users can interact to accurate the topic more and more.

5. Chronological Visualization

The chronological visualization illustrates stories that are clustered by similarity each other, arranged along time axis as a topics (see also **Fig. 3**). The similarity among stories calculates nearness of the time and contents. In Fig. 3, dashed ovals group the similar stories. However, stories are not separated clearly. Therefore, actual visualizing system will not display the ovals.

Users are able to view stories on the interface to click the story nodes. The interface enables to accurate the topic in their mind. For traditional method of searching, it prepares input form for word queries. The other method provides interactive interface on the visualizing part. Every story on the visual interface are selectable. The searcher make query internally and put into the interface when the user selects them.

6. Conclusion

In this paper, we presented a concept of representation time-dependent events in the Web. We gave the definitions of the terms of story, event and topic. It is necessary to illustrate Web documents as timeline to retrieve stories through a time. We also presented the chronological visual interface to interact with users. The interface showed stories in a timeline, and the visualization in the interface showed clusters of stories that were similar mutually.

We will realize this interface and the back end modules. We will evaluate availability of this system experimentally.

References

1) Allan, J., Papka, R. and Lavrenko, V.: *On-line New Event Detection and Tracking*, Proc. of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 37–45 (1998).

2) Brin, S. and Page, L.: *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Computer Networks and ISDN Systems, vol. 30, pp. 107–117, (1998).

3) Donath, J., Karahalios, K. and Viegas, F.: *Visualizing Conversations*, Proc. of the 32nd Hawaii International Conference on System Sciences (HICSS-32) (1999).

4) Fertig, S., Freeman, E. and Gelernter, D.: *Lifestreams: An Alternative to the Desktop Metaphor*, Proc. of ACM SIGCHI Conference on Human Factors in Computing Systems Conference Companion (CHI ’96), pp. 410–411, ACM Press (1996).

5) Nanno, T., Suzuki, Y., Fujiki, T. and Okumura, M.: *Automatic Collection and Monitoring of Japanese Weblogs*, WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2004).

6) Tsukada, K., Takabayashi, S. and Masui, T.: *Dying Link*, Proc. of the 10th International Conference on Human-Computer Interaction (HCI 2003), Vol. 3 (Human-Central Computing), pp. 1353–1357 (2003).

7) Rekimoto, J.: *Time-Machine Computing: A Time-centric Approach for the Information Environment*, ACM UIST’99 (1999).

8) Ringel, M., Cutrell, E., Dumais, S. and Horvitz, E.: *Milestones in Time: The Value of Landmarks in Retrieving Information from Personal Stores*, Proc. of Interact 2003, pp. 184–191 (2003).

9) Kumar, V., Furuta, R. and Allen, R.: *Metadata Visualization for Digital Libraries: Interactive Timeline Editing and Review*, Proc. of DL 1998, pp. 126–133 (1998).