A Browser for Browsing the Past Web

Adam Jatowt[1,2], Yukiko Kawai[1], Satoshi Nakamura[1], Yutaka Kidawara[1] and Katsumi Tanaka[1,2]

[1]National Institute of Information and Communications Technology
3-5 Hikaridai, Seikacho, Sorakugun,
619-0289 Kyoto, Japan
Phone: +81-77-498-6828
{adam, yukiko, gon, kidawara}@nict.go.jp

[2]Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku,
606-8501 Kyoto, Japan
Phone: +81-75-753-5969
ktanaka@i.kyoto-u.ac.jp

ABSTRACT
We describe a browser for the past web. It can retrieve data from multiple past web resources and features a passive browsing style based on change detection and presentation. The browser shows past pages one by one along a time line. The parts that were changed between consecutive page versions are animated to reflect their deletion or insertion, thereby drawing the user's attention to them. The browser enables automatic skipping of changeless periods and filtered browsing based on user specified query.

Categories and Subject Descriptors
H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia – *navigation*

General Terms: Algorithms

Keywords: Past web, web archive browsing, web archives

1. INTRODUCTION
Archiving the web started several years ago with the aim of preserving previous web pages for future use. There are general web archives containing any kinds of pages, such as Internet Archive [3] and dedicated past web repositories. Topical archives (e.g., [1,5]) preserve the history of certain topics or events, while country-specific archives preserve web documents coming from a particular country (e.g., [4]) or a group of countries [2]. Other repositories of past web pages, such as local caches, site archives, transaction-time servers, and search engine caches, can also be considered to some extent as web archives. Web archives store snapshots of the web as it was in the past and contain a vast amount of data that can be used for many purposes. Mining information from the past web can help in better understanding the past and, more importantly, understanding the present. Past content of the web could be also potentially used to predict the future or can simply have sentimental value.

By the past web, we mean here the set of all pages that were ever available on the web in the past. Past versions of pages are their "frozen" snapshots stored in web archives. The past web, like the current web, can be browsed spatially by following links between past page versions. However, additionally, pages can be browsed chronologically by viewing their snapshots over time. We call these browsing styles vertical and horizontal, respectively.

In this paper, we describe a past web browser that supports vertical and horizontal navigation of the spatio-temporal structure of the past web. The proposed browser uses meta-archive approach for extracting data from several past web repositories to reconstruct web page histories. Due to resource limitations, web archives generally contain incomplete sets of previous versions of pages, or they do not have any for some pages at all. Thus, representation of previous page states is often poor and incomplete. Combining data extracted from multiple sources of web snapshots can improve the accuracy of page histories. The client-side work mode of the browser enables customized and passive viewing of past web pages. This is not possible while browsing web archives using their proprietary interfaces [2,3]. The browser supports horizontal browsing of the past web by change detection and presentation enabling quick recognition of the changing content and viewing the evolution of pages over time. This allows also the periods without change to be skipped, which speeds up browsing and focuses the user's attention on the new content. Additionally, the browser is equipped with search-like functions that enable filtering of the displayed changes based on the user's query.

In the next section we describe the proposed browser, and in Section 3 we describe the implementation. We conclude in Section 4 with a brief summary.

2. BROWSER
After receiving a request from the user for a page version at a certain point in time in the past, the browser communicates with its network of web archives, querying them about the page versions they contain. The archives should then send lists of the page versions they contain, ideally with information about those with changed content. The browser downloads the version closest to the specified point in time, plus next, selected versions. The browser attempts to download mostly those versions that contain different content to minimize cost, yet, at the same time, to accurately reflect all changes. The received pages are ordered based on their timestamps, which are provided by the past web resources. This meta-archive approach brings the user closer to the experience of browsing the past web rather than browsing individual web archives. However, since web archive interfaces are not standardized, the browser must currently use different data acquisition methods when communicating with the collaborating resources.

During horizontal browsing past page versions are presented to user one by one, like frames in a movie. The content differences (additions and deletions) between consecutive versions are detected and animated to catch the user's attention. In this way the user can see changes in the past content of the page. The static (unchanged) content remains the same, except that it can shift

Copyright is held by the author/owner(s).
WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

position. Pages are processed from top to bottom, meaning that changes at the top are shown first. The deleted content first blinks for some time and then disappears while added content first appears, then blinks and finally remains on the page. Both contents are also marked by different colors. In this way, pages are processed one by one. After the processing of each page version is completed, the system pauses to enable the user to quickly view the whole page. Then the next page version is presented. The user can control the speed and direction of browsing with a speed meter. The list of available page versions with their dates is also shown together with a timeline view to enable users to easily navigate during horizontal browsing. The timeline additionally shows the distribution of changes over time. The user can click on a certain page version in the list or on the timeline to jump in time. Both the list and timeline track the position of the currently viewed page version. The passive style of page history viewing requires minimal user interaction while clearly showing the evolution of the page content over time.

The user can stop the horizontal browsing at any time by pressing stop or pause buttons, similar to those on video players. He or she can then view the currently displayed page version in detail or follow a link on the page. If the user clicks on a link, the browser loads the version of the linked page closest in time to the page being viewed and continues the presentation from that page version. The user can return at any time by pressing one of two back buttons. The page-version-consistent back button reloads the page version containing the link that was followed while the time-consistent back button loads the version of the previously viewed page closest in time to the currently displayed page.

As mentioned above, there is a manual jump capability. It enables the user to input a new date or URL or to click on any page version on the timeline. Additionally, the browser features also an automatic jump facility (Figure 1), which improves horizontal browsing in case of the abundance of static, redundant content in past page versions. During the presentation flow, the browser jumps over the changeless periods and displays the first page containing any changed content. Automatic jumping can be switched on or off by the user. This function speeds up and improves the browsing experience thanks to limiting the displayed page versions to the ones that contain changes.

Finally, the browser is also equipped with search-like function for filtered browsing. If the user enters a query, only changes containing the query words are shown with animation. The rest of the page is treated as static content and is not animated. In this way the user can view the historical content of the page that is related to his interest.

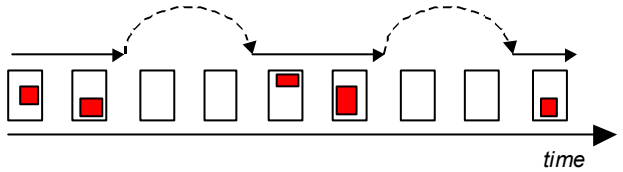

Figure 1 Automatic jumping (dashed lines) during horizontal browsing; changed parts are shown in red

3. IMPLEMENTATION

We built a prototype of the proposed browser in C#; its interface is shown in Figure 2. Past versions of web pages are obtained from Internet Archive, Google and MSN search engine caches, and from local cache. Changes are detected by using the diff algorithm for consecutive versions of pages after the HTML nodes on each version are converted to single lines. The changed nodes are marked by <del> and <ins> tags corresponding to the type of change. The deletions and insertions are animated by interchangeably switching the visibility styles of the HTML nodes between "visibility: hidden" and "visibility: visible" in order to obtain a blinking effect. After animation the content bracketed by <del> tags is set to "display: none" to finally hide the node and accommodate its freed space. The content bracketed by <ins> is set to "display: inline".

Caching and link prefetching minimize latency. During horizontal browsing, up to some number of consecutive page versions are fetched so that the system can later retrieve their data from local cache. If the pause or stop button is pressed, the browser starts the link prefetching process to retrieve lists of available versions of links occurring on the currently viewed page version.

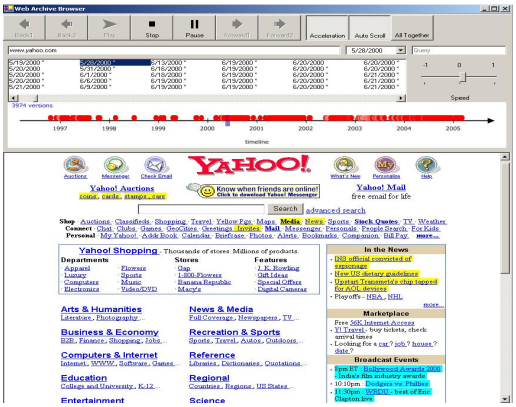

Figure 2 Interface of past web browser

4. CONCLUSION

In this paper we have proposed a browsing interface to the past web. The browser can collect page versions from multiple past web repositories in real time, thereby improving the accuracy of the representation of the past web. It features a customized, passive browsing style that incorporates animated presentation of detected changes. Additionally, it allows for change-oriented and filtered browsing.

5. REFERENCES

[1] Election Web 2002 Archive: http://lcweb4.loc.gov/elect2002

[2] Hallgrimsson, T. and Bang S.: "Nordic Web Archive" 3rd ECDL Workshop on Web Archives in conjunction with 7th European Conference on Research and Advanced Technologies in Digital Archives, Trondheim, Norway, 2003

[3] Internet Archive: http://www.archive.org

[4] Pandora, Australia's Web Archive: http://pandora.nla.gov.au

[5] September 11 Web Archive: http://september11.archive.org