What Can History Tell Us? Towards Different Models of Interaction with Document Histories

Adam Jatowt[1], Yukiko Kawai[2], Hiroaki Ohshima[1] and Katsumi Tanaka[1]

[1]Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan
Phone: +81-75-7535969
{adam, ohshima, tanaka}@dl.kuis.kyoto-u.ac.jp

[2]Kyoto Sangyo University
Motoyama, Kamigamo, Kita-Ku
603-8555 Kyoto, Japan
Phone: +81-75-7052958
kawai@cc.kyoto-su.ac.jp

ABSTRACT
The current Web is a dynamic collection where little effort is made to version pages or to enable users to access historical data. As a consequence, they generally do not have sufficient temporal support when browsing the Web. However, we think that there are many benefits to be obtained from integrating documents with their histories. For example, a document's history can enable us to travel back through time to establish its trustworthiness. This paper discusses the possible types of interactions that users could have with document histories and it presents several examples of systems that we have implemented for utilizing this historical data. To support our view, we present the results of an online survey conducted with the objective of investigating user needs for temporal support on the Web. Although the results indicated quite low use of Web archives by users, they simultaneously emphasized their considerable interest in page histories.

Categories and Subject Descriptors
H.5.4 [**Information Interfaces and Presentation (e.g., HCI)**]: Hypertext/Hypermedia

General Terms
Algorithms, Theory

Keywords
document history, past web, time travel, archiving, versioning

1. INTRODUCTION
Versioning was an intrinsic part of many early hypertext systems. For example, it was a fundamental mechanism in Nelson's Xanadu [16] used there in order to prevent content from being duplicated between different versions of documents. Instead of copying the same parts of content, newer versions of documents contained links to the content of older versions. However, versioning and version management do not currently seem to play a major role on the Web. While some sites provide a browse or search access to their own archives, many times past versions are often seen as a kind of by-product of document evolution, and retained for security reasons if kept at all. In general, with some exceptions such as multi-authored documents, past document versions are usually not visible to the public. Thus, the Web is in fact a dynamic medium where few are concerned about the previous content. Some effort has been made by the Web-archiving community toward preserving the Web; however, an average Web user usually has no idea on where to look for previous versions of documents or on how to use them.

This paper proposes extending the interaction with Web pages by empowering users to access and analyze the historical data of documents. We have categorized different interaction mechanisms that could make this historical data useful. Tighter integration of documents with their past content could bring about better understanding of a document's evolution and help users to obtain the context of the current version, and, in general, document's long-term topics and characteristics. For example, recent topic drifts or a document's general frequency of change could be estimated. This should improve the perceived trustworthiness of the document. Another advantage is temporal search and the facilitated re-discovery of past content. We discuss several potential ways of adding historical components to documents and empowering users with the freedom and means to use past Web data. We then demonstrate applications that we have designed to realize the proposed interactions.

The remaining part of the paper presents the results of a questionnaire that was prepared for a large group of users to provide more insights into possible interaction types they could have with document histories. To the best of our knowledge, this is the first such kind of study related to access and use of past data on the Web. Lastly, we provide a discussion on various issues and obstacles related to our proposed models of interaction.

In conclusion, the contributions we make in this paper are three-fold:

- We attempt to raise awareness and the importance of providing historical context for changing documents.
- We list several potential history-based interactions by users and propose systems supporting such interactions.
- We present the results of a survey we conducted to understand potential use cases and the attitudes of users with respect to temporal support on the Web.

The remainder of this paper is organized as follows. The next section provides the background and Section 3 discusses related research. Section 4 contains our proposals for history-based

HT'08, June 19–21, 2008, Pittsburgh, Pennsylvania, USA.

interaction mechanisms. Section 5 presents the results of a user study that we carried out. Section 6 contains a discussion on various issues related to the proposed approach. The last section concludes the paper.

2. BACKGROUND

In this paper we have assumed the existence of past data on document histories. There are basically two ways in which document histories are preserved. Historical data can be preserved by a document's author (or owner) or it can be archived by external, specialized institutions. The former is often triggered by document updates and is usually done on individual sites or pages. In contrast, the latter is basically independent of document update patterns and is often done on larger collections of documents. As a result, some historical content is lost (or the same content is re-crawled). This inability to capture the complete evolution of documents is also aggravated by the typically large size of archived collections.

2.1 Web Versioning and Archiving

The main purpose of versioning is usually to enforce rigorous tracking of sources so that consecutive changes and their authors can easily be identified [6, 20]. This guarantees a rollback option for any previous version of a document. However, only certain types of pages (e.g., wikis) on the current Web enable strict versioning and source tracking. The reason for versioning these pages is often related to the development process in creating documents. For example in collaborative environments (e.g., Wikipedia pages) versions are often seen as subsequent stages of the generation of a document. Storing the past versions guarantees that the previous state of the document can easily be restored when mistakes or collaboration failures occur. The development process implies that consecutive versions of the document should be better than those previous along with the process of the document's development. The travel back through time for such documents would basically only be useful for observing how the document evolved, i.e., looking at its stages of development at a given point in time or just restoring the previous versions.

This is, however, not necessarily true for the majority of Web documents. Usually, the reason content on the Web is modified is basically due to alterations to the various contexts of documents (e.g., changes in authors' ideas and facts); thus, they can simply be triggered by the time flow rather than by the document's improvement and development processes. For example, certain information ceases to be valid and is thus erased while other content appears. Or, in blogs the passage of time makes authors produce new content to substitute for the old one. This kind of adaptation of a document to its changing contexts makes traveling back through time an interesting and useful experience that can tell much about the document in general, the role it played in the past, or the underlying real-world objects represented by it.

There have been some proposals to automate version management on the Web. The DeltaV protocol[1] is an extension to the Web Distributed Authoring and Versioning (WebDAV) protocol, for supporting remote versioning and configuration management of documents stored on Web servers. WebDAV, on the other hand, was designed to extend Hypertext Transfer Protocol (HTTP) in order to enable remote collaborative authoring of documents. Dyreson et al. [3] proposed transaction-time Web servers that would automatically preserve previous versions of documents in all content updates and provide standardized methods of accessing these.

However, since, generally, version management is not currently provided for the majority of pages, external, third-party data sources such as Web archives need to be exploited if one wishes to interact with document histories. Along with the dramatic growth of the Web, various archiving institutions began large-scale-preservation projects to save the Web for future generations. The Web thus started to be seen as the part of humanity's heritage. The Internet Archive[2] [14] is the best-known public Web archive containing more than 2 petabytes of data composed of periodical crawls delivered from the Alexa [3] search engine. Wayback Machine is an interface to access the Internet Archive's page collection. It provides results in the form of a directory page that lists links to page snapshots contained in the repository. Other Web archives also exist such as those focusing on certain countries (e.g., the Australian Archive[4]) or thematic Web archives (e.g., the September 11 archive[5]). There are also other repositories of past Web data such as local caches, site archives, personal repositories, or search-engine caches. For interested readers, Masanes [14] provides a thorough description of many issues related to Web archiving.

2.2 Different Types of Time in Pages

We represented page histories longitudinally in this research. Each document was assigned a unique ID (i.e., URL) at any point in time. The history of the document was linear (without branches) and was represented by a series of past snapshots taken at certain points in time in the past. Although, we only focused on a single document, the larger neighborhood of a document could also be considered here.

We distinguished four different times related to Web content. The first was the time for the content to occur on pages. This time is analogous to the transaction time in databases, which specifies the period when content is stored in database records (here, a Web page). It is similar to the notion of *constructive time* proposed by Shipman and Hsieh [18].

The second time was the time for content to become valid. This notion of time is related to real-world entities and events. For example, information about hiring new employees on a company's homepage is valid as long as the company is actually seeking new staff. This time can be estimated by detecting and anchoring temporal expressions occurring on page content. Naturally, the transaction and valid times may differ to some extent. For example, the information on job opportunities may still appear on the page despite the positions having already been filled. However, users often automatically assume a correspondence between both these times.

The third time is the interaction period a user has with the document. This embraces any time points when he or she accesses the page and undertakes any activities on it. This time corresponds to the *interactive time* proposed by Luesebrink [13].

[1] http://www.webdav.org/deltav

[2] http://www.archive.org

[3] http://www.alexa.com

[4] http://pandora.nla.gov.au

[5] http://september11.archive.org

Finally, the social time is the fourth time when other users interact with the same document.

The interplay between the above times should provide opportunities for establishing new types of interactions and building applications to support these interactions. We will discuss some of them further in Section 4.

3. RELATED RESEARCH

Shipman and Hsieh [18] proposed the history navigation of hypertexts on an example system called the *Visual Knowledge Builder* designed for organizing and interpreting information. Their objective was to allow readers to observe and understand how the hypertexts were developed, author writing styles, and the general context of documents. The system allowed automatically replaying a document's history with a chosen speed.

Francisco-Revilla et al. [4] studied how users perceive changes in Web pages as part of a project called *Walden Paths*, which investigated ways of managing collections of Web resources for educational purposes. The authors identified several key aspects that affect the importance and usefulness of changes perceived by users.

The International Internet Preservation Consortium [7] provided an exhaustive list of particular cases in which Web archives could turn out to be useful for users in completing certain real-world tasks. There was however, no empirical investigation into the selected cases. Here, we propose more general history-based interaction mechanisms for users browsing the Web and discuss challenges with these and their potential.

Wexelblat and Maes [19] demonstrated the *Footprints* system that utilizes historical data on user visits to documents by adding a novel social context to different browsed structures. Their objective was to guide new users to useful and popular resources. These ideas later formed the basis for research devoted to social navigation and social searches (e.g., see Freyne et al. [5]).

McCown et al. [15] have recently measured the persistence and availability of page copies in search engine repositories and the Internet Archive. This analysis was part of a project called *Warrick* that was aimed at helping users to reproduce the latest content from their Web sites when Web data were lost due to various mishaps such as server crashes.

Luesebrink [13] distinguished various times specific to hypertext literature. For example, the interface time is the time span for the reader to interact with the document, while the cognitive time determines the chronological ordering of events in the narrative. We would like to emphasize that document versioning and version management are not our main foci here. There is already a large body of research devoted to these issues (e.g., see Vitali [20]). In contrast, we have concentrated on potential ways in which document history can support users' activities on the Web.

4. INTERACTION WITH DOCUMENT HISTORY

In the real world, we often look at the histories of various objects such as companies, institutions, countries, and people. These histories tell us a great deal about the previous states of the objects, their general characteristics, their current states, and even future perspectives. In the same way, we can analyze the lifetimes of documents by looking at their past.

In this section, we seek to establish certain links to past data depending on user needs. These links may not only lead to particular versions of past documents but to summary-like views of page histories. They can also be made dependent on personal interactions of users with the content over time.

4.1 Access to Past Page Snapshots

Here, we talk about access to a document's history to view content the document had at a certain point in time in the past. This is standard way of access, which is usually enabled by most Web archives. Users may access page histories in this way for a variety of reasons. For example, they may wish to revisit some content from the past that they have seen before but that no longer exists in the document. They may also merely wish to check what was published on the page before, e.g., to search for interesting content in addition to that already shown on the version of the present page. Occasionally, users may also not be able to access the current versions of documents due to server or connection problems. In such cases, they may want to view at least the last saved versions of the documents. This actually seems to be why "cached" links are often added to search-engine results that point to recent snapshots of documents stored by these engines. We propose inserting such links automatically within the browser.

Users could access past content of Web documents through proprietary interfaces of Web archives. This access is usually URL-based and a given point in time has to be chosen. If a specific time point for access is not known, the user is then left to explore available past snapshots that may actually be burdensome where many page snapshots have been preserved. In general, there is rather limited support for users to access document histories and existing interfaces do not make it easy for users to search or browse page histories.

We believe that the access points to the history of a given page should be immediately available to users once they visit it. Searching for the location of available past snapshots of the page and aggregating them so that they can be presented to users as potential targets should be automatically done by Web browsers. Such temporal support, apart from relieving users of burdens related to reconstructing page histories, would also most likely encourage them to analyze the history of pages.

Figure 1 outlines the concept underlying the temporally supported model of Web browsing where users browse pages as they can currently do using standard browsers. However, from time to time, when needed, they can make jumps into a document's history to view certain content the document had in the past or even browse its history longitudinally to observe its evolution. It may also be possible to click links on past copies of pages (Fig. 1) and view the content of linked pages around that time, provided that the data is accessible [6]. In this way, past Web pages can simultaneously be browsed while browsing the present Web in the usual manner. Standard browsing of the present Web would be a major type of browsing, while traveling back through time is an optional means of providing complimentary information on visited pages.

[6] The Internet Archive rewrites links inside page snapshots stored in its repository to direct them to copies of corresponding pages that it also contains.

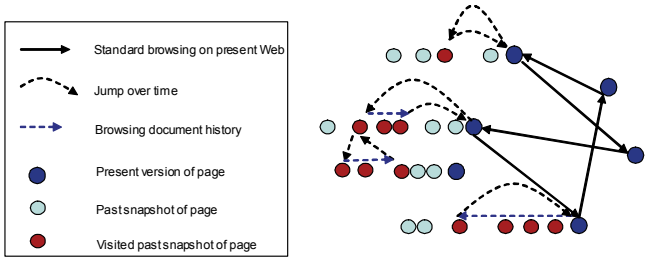

Fig. 1 Temporally supported model of browsing.

4.1.1 Application Example

To facilitate travel back through time in the above browsing model, we propose an application called the Past Web Browser [9] that merges past snapshots from different collections, providing a kind of virtual link from the present content of documents to their past snapshots and, at the same time, adding some browsing functionalities. It provides an enhanced model of navigation and browsing that is similar to some extent to slideshows or VCR players. The system helps users to browse document histories through a passive (as a slideshow of consecutive page snapshots) presentation mode along the time axis, automatically presenting consecutive versions of pages, emphasizing changes over time, and skipping history periods without any changes to content. It also allows for more restricted browsing by emphasizing content changes that contain user-specified terms. When a user issues query-like terms, the browser then automatically presents only those content changes that contain a given set of terms skipping the remaining parts of the history. Figure 2 is a screenshot of the interface for the Past Web Browser.

The above-discussed browsing model has been to some degree implemented in *Waxtoolbar* [7], which is a Firefox browser extension for providing access points to the Internet Archive repository. In contrast, Past Web Browser, offers more functions for navigation and browsing as well as it can retrieve past data on pages from other sources.

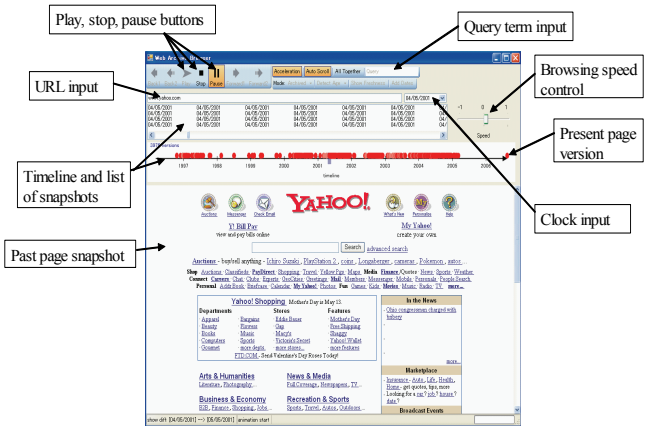

Fig. 2 Interface of Past Web Browser.

4.2 History Summary

On numerous occasions, users do not wish to find particular information or access particular past versions of pages. They may also be hesitant about browsing through document histories as the number of past document snapshots may be excessive. In correspondence to the real world, as Wexelblat and Maes pointed out [19], people often prefer to receive a brief summary of a meeting rather than to listen to the entire recording. In such cases, they may rather want to obtain a general idea and an overview of the page history. The document history may thus be summarized and presented concisely so that it can be more easily digested by users (Fig. 3). The document's summary here can be considered as a kind of short "biography" of the document describing its entire lifetime or a selected part of it.

One reason the historical summary of a document would be useful is in helping users to better understand the document and its characteristics. In a real-world situation, changeable communication media such as printed magazines or newspapers are described by their themes or topical scopes. For example, magazines can be devoted to such topics as soccer, financial markets, and engineering. Describing media by using their long-term topics enables us to capture their inherent themes and characteristics and helps users to choose documents that overlap with their interests. In a similar fashion, online documents can be characterized by analyzing their past content.

To sum up, summarizing document histories has the potential for improving our understanding of them through providing answers about their long-term topics or long-term document characteristics. This can also enhance the documents' trustworthiness as there is more empirical information provided on the characteristics of the documents.

History summary can be in a form of simple statistics derived from available page snapshots. For example, users may be informed about the average frequency of changes, the degree of change, or the age of a document. A summary of the page history can also be in the form of free text extracted from historical content. For example, Jatowt and Ishizuka [8] detected salient terms from the page history and later used these for extracting representative sentences.

An interesting use of a document summary would be to predict future page content. For example, suppose a system detects that submission deadlines have been extended for several consecutive years for the same series of conferences. It is then quite probable that the deadline for submitting papers to the upcoming conference is also going to be extended next year. This example is actually related to the periodic characteristics of a real-world object (a conference in the above case), whose information is published online. The summary of the historical data of pages related to that real-world object can then be extrapolated to be considered as the summary of the characteristics of this object to some extent. Users could, for example, learn about typical topics at these conferences and their changes or drifts over time, or find information related to changes in program or steering committees provided there had been any in the past. The above example is also indicative of the problems related to tracking pages over time. Some objects may have different URLs during different time periods. Conference Websites may often have different URLs each year, companies may change their names and the URLs of their homepages, or, simply, Web sites can be restructured.

Historical summaries can also be tailored to specific needs and the ways of creating historical summaries of documents should be dependent on the goals of users and the role that these summaries

[7] http://archive-access.sourceforge.net/projects/waxtoolbar

will play. For example, they may be query-dependent, time period dependent, or user-dependent.

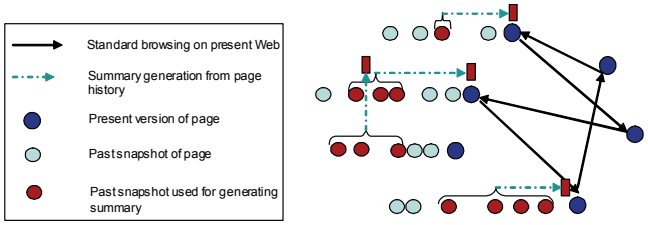

Fig. 3 Browsing with created summaries of page history.

4.2.1 Application Example

Below, we briefly discuss an example of a tool for analyzing histories called a Page History Explorer [11], which is capable of generating short visual summaries of document histories (Fig. 4). Users can decide the URL of a requested page and the time period for analysis. The system then searches for available past snapshots of the document that were crawled within a specified period of time. These are then sampled and used to provide a summary overview of page content and outlook within that time period. The outlook of the page is displayed as a series of thumbnail images of document snapshots arranged chronologically in the horizontal dimension. The vertical arrangement indicates the difference in content between consecutive page snapshots. The further apart the neighboring snapshots, the larger their difference in content. The system also represents the document's past content as a term cloud where the size of a term indicates its prevalence in the document history. Thus, if a given term has occurred often in the past content of the page, then it will be in large font and be immediately visible to users. Term clouds can also be built using the activity levels of terms. In such cases, terms that are often added or removed from the page throughout a given period of time are deemed to be active and shown to users. This helps them to characterize the average change in content on the page. In addition, the Page History Explorer gives the series of unit-term clouds for shorter time segments to the main-term cloud, which characterizes the whole time period specified by the user.

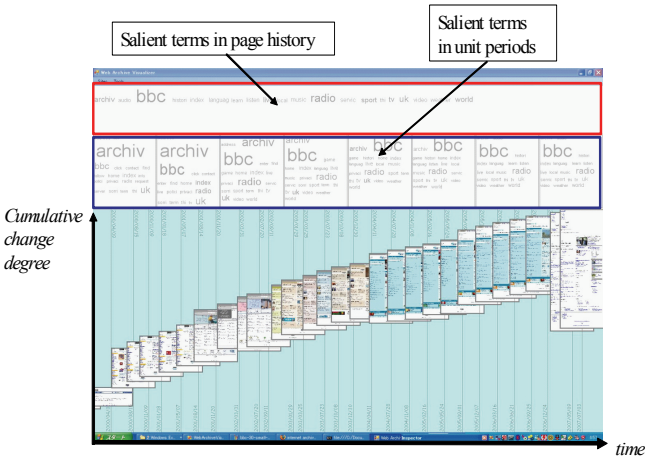

Fig. 4 Example of historical summary of page.

4.3 Enriching Page Content with History-derived Information

Past page content in its original or summarized forms can be contrasted with the current version of the page when browsing or during other activities that users intend to do. In such cases, we can provide a kind of temporal context for documents.

Simply contrasting pages and their temporal contexts may, however, not be really effective. We think that users can become interested in mapping the information derived from a document's history on the current document. For example, those that change most often or, in contrast, the most-static document areas can be identified and indicated on the page. In other words, we sought methods of finding a connection between a page's present version and its history and we wished to modify the view of the present document to demonstrate this connection to users.

4.3.1 Application Example

Below, we present an example of how a document's history can be mapped on the current-page content to enhance a user's browsing experience. We have built a tool for detecting the age of document elements and for indicating information about it on the document to be displayed to users [10]. They can thus see how old the particular elements of the document are or what the average age of its content is.

Changing documents may often contain content elements introduced at certain points in time. However, for a visiting user these elements look as if they were created at the same time. Of course, document content is occasionally annotated with the time it was originally created. For example, blogs have timestamps indicating the age of their posts or news articles have dates attached that indicate their age. These annotations provide necessary context for users and often change the users' perception of content elements. In many cases, however, there are no clues as to the age of a document's content. The reason for this is that authors did not see the need for adding such information or they may have forgotten to do so. However, users may wonder how old certain content elements are for a variety of reasons, especially in the case of time-sensitive content. For example, an investor may wish to know for what length of time a certain statement has appeared on a company's page. Or, there may be annotations for enhancing traffic to page content purposely written by page authors such as expressions like "new" or "recent". The trustworthiness of such temporal annotations can thus be checked by automatically analyzing document histories.

Our proposed approach is to search in document histories for the oldest page snapshots that contain the same content elements as the current version of the page. The timestamps on these snapshots allow the approximate dates of content appearing on pages to be estimated. The results are later annotated on the page within red frames indicating content of the same age. The detected creation dates of content are attached to the bottom right corners of the frames (Fig. 5). In this way, the oldest and newest content areas can be identified and the content can be chronologically ordered on pages. The precision of the process for estimating age depends on the amount of past data that is available. Also, having estimated the age of content elements on a page one can calculate the average age of the content on the entire page to have a general idea of its currency.

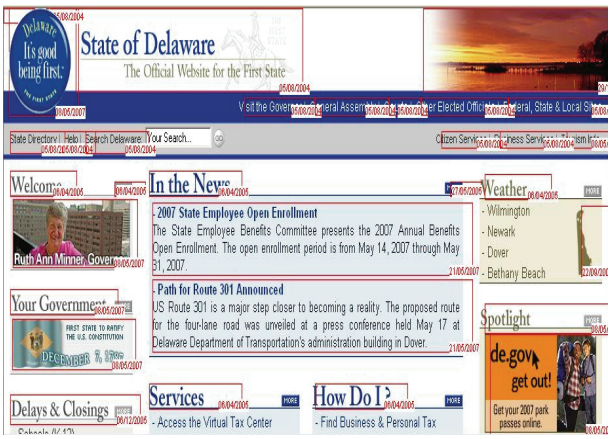

Fig. 5 Example of page annotated with content-creation dates.

4.4 User Interaction History

4.4.1 Revisiting Support

This section explains how past information on user revisits to a page may be used to improve his or her browsing experience on the current Web. Here, we focus on the interplay between the time of user interactions and the transaction time of Web content. Users frequently access the same pages and revisits are generally quite common on the Web [17]. By simply detecting and indicating new content in documents it is possible for them to understand page transitions that have occurred since they last accessed these pages. This is especially useful for documents that undergo a few or numerous changes. According to the concept of change blindness, people often do not notice what was changed in documents they visited.

A simple way of facilitating awareness of change in re-visited pages is to detect and mark changed content on documents [12]. This is done by utilizing the user's browsing history. In addition, changes in the neighborhood of a document are visualized in a way that is seamlessly incorporated into browsing. That is, all the links pointing to pages already visited by the user are annotated with how current their content is from the user's viewpoint. The annotations are in the form of numbers added next to any link indicating the ratio of content on the target page that is new to the user. The color of the link is changed on a scale from blue to red where blue denotes completely fresh content while dark red indicates known (already visited) content. In this way, we can adapt the customary color-coding of links to indicate changed areas in the neighborhood of the document and to direct users to novel content. Users can thus follow links that contain much fresh content when they search for new information. This should prevent them from wasting time and resources enabling browsing and navigation that is driven by freshness.

Note that other tools may also support revisiting users in certain ways, e.g., there is a large body of research devoted to various aspects of adaptation in hypertext documents (e.g., see Brusilovsky et al. [2]).

4.4.2 Page Popularity over Time

When accessing documents users usually have no information about their popularity or previous interactions other users have had with the document. In contrast, as pointed out by Wexelblat and Maes [19], real-world objects often contain evidence of their histories, e.g., stairs or monuments are worn out by being frequently used and touched. Social navigation has recently become an active field of research where traits of user activities on documents are utilized to guide new users to the most popular or the most important content. For example, Freyne et al. [5] demonstrated how community-derived social information could support search and navigation in digital collections. Their method also allowed resources to be selected that had most recently been accessed by other users.

However, it is usually difficult to know about the popularity of Web documents and their evolution without having document-access logs. On the other hand, social bookmarking, which has recently gained a great deal of attention as a popular Web service, can provide necessary data on temporal changes in the popularity of documents [21]. Social bookmarks are usually annotated with timestamps indicating when they were created. We propose integrating page histories with the information on the evolution of their social bookmarks. By analyzing the number of bookmarks a given page has received over time it would be possible to view the changes in the popularity of the page over time. In this way, historical interaction by users with the page could be utilized. Time periods when pages are increasingly bookmarked denote phases when the pages were deemed to be popular. This kind of information might indicate interesting points in time whose historical content should be visited or should be used to construct temporal summaries. Of course, we need to note that not all users bookmark all pages they visit and some pages only have a few social bookmarks or do not have them at all.

4.5 Historical Comparison

Users in the real world frequently want to compare histories of different objects. Comparisons of document histories can be done between different versions of the same document, versions of two different documents captured around the same or different times, or between the historical summaries of these documents. In addition, if we annotated parts of the content of documents with their creation dates, as in the screenshot in Fig. 5, we would know which document provided the latest content. It would thus be possible to implement freshness-centric searches for constrained document collections in which the documents that contained the most recent content related to specified keywords would be returned to users.

5. USER STUDIES

Here, we report on the results of a survey we did regarding temporal aspects of the Web and user needs for gaining access to document histories. We asked subjects several questions to analyze their attitudes to historical content on the Web and their willingness to interact with page histories. The questionnaire was administered from the 7th to the 12th of February 2008 to a group of 1000 Internet users in Japan. The subjects were divided into four groups depending on their ages: 20-29, 30-39, 40-49 and 50-59 years old. Each group consisted of 250 respondents, where half were males and half were females. Each respondent received a small amount of money to complete the survey. The survey was done in Japanese (the results were translated). The findings obtained from questions related to historical data on the Web are discussed below.

Figure 6 shows the results of analysis aimed at finding out how many subjects used Web archives in comparison to other Web

content. The percentage of subjects that used any Web archives, such as the Internet Archive, at least once a month is actually very low. Only 19 subjects responded that they had done so (1.9%). To some extent, the reason for this may be the lack of large Web archives open to the public in Japan. Many subjects also did not seem to be aware of the existence of Web archives. Actually, during the course of our studies we found that many people were often quite surprised to learn about the existence of repositories preserving large portions of historical Web content. They also often expressed enthusiasm on hearing about the possibility of freely accessing such content. The limited access and lack of effective temporal-search functions may have been other reasons for the lack of popularity of Web archives.

Figure 7 shows the answers to questions on temporal context that subjects would like to have when encountering information on the Web. There were seven potential answers provided and the users had to choose three that were most important to them. We can see that the currency of information is the most important temporal factor that the users paid attention to; 66.8% of the respondents selected it as their first choice. The age of content on a page was another aspect frequently chosen (12.8% of participants selected it as their first choice, while 26.5% selected it as their second choice and 23.5% as their third). The popularity of information on the Web and its evolution were less important to users.

Our next study focused on page content. We asked subjects what sort of information they would like to obtain if they could access page histories (Fig. 8). They had to choose three answers from a total of seven. We asked whether they would like to know the ages of pages, sites, and certain content elements on the page, including a summary of past content and information on the parts of content that had changed since the users last visited the page. All answers were generally chosen quite frequently. The most popular choice was information about the age of the site and the age of the page (34.2% of participants selected the former and 21.1% selected the latter as their first choice).

Figure 9 shows the answers to another question about access to past content of pages. The subjects could select the top three answers from a total of six. Three answers had a personal aspect: a) to view content that they had already seen (but was no longer accessible on the page), b) to view content that they could not see (e.g., due to limited Web access) and c) to view content that was inaccessible on the hosting Web server (e.g., due to server problems). The other three answers were: d) to check how old the page was, e) to see how the page had changed in the past, and f) others. Answers that did not have any personal aspects were selected less frequently than those featuring personal aspects. The top two answers were: wanting to revisit content that had already disappeared and to view content that could not previously be accessed (49.4% of participants selected the former and 29.2% selected the latter as their first choice).

Our last question concerned the types of pages for which subjects would have liked to view their histories (Fig. 10). Here, the respondents were asked to select five answers out of a total of nine. The histories of news sites were selected most often as the first answer (42% of participants). This was not surprising as news articles contain time-dependent content. Subjects also would have liked to view the histories of pages related to their interests and hobbies (30.7% of respondents chose this as their first answer). Histories of pages related to residences or locations where they had worked (7.8%) or schools they had attended (2.5%) were definitely less interesting. These results, to some extent, imply that the subjects were interested in histories of pages with time-sensitive content and in pages with content that was related to their interests. These findings may perhaps influence the selection process for content to be preserved in Web archives. Of course, archivists, historians or other professionals may have different requirements and needs regarding the types of documents to be archived.

We also checked whether participants were interested in changes to results retrieved by search engines. Some 14% of subjects admitted that they occasionally issued the same query to search engines at different times to check what had actually changed on the Web. For example, they may have checked whether there was any new content related to their hobbies or any new pages mentioning their names or favorite movies. We concluded that there should be another kind of query – comparative-informational query that could be included in the well-known query taxonomy discussed by Broder [1] (i.e., informational, navigational, and transactional queries). This finding can be utilized to design systems supporting the visualization of changes in search results over time.

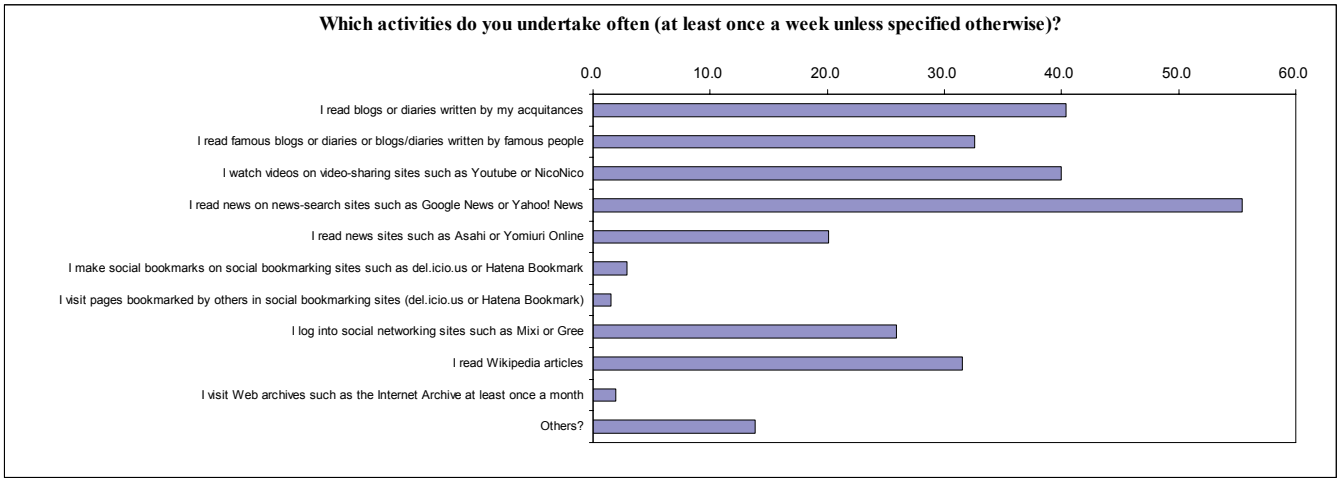

Fig. 6 Questions on frequent user activities on Web.

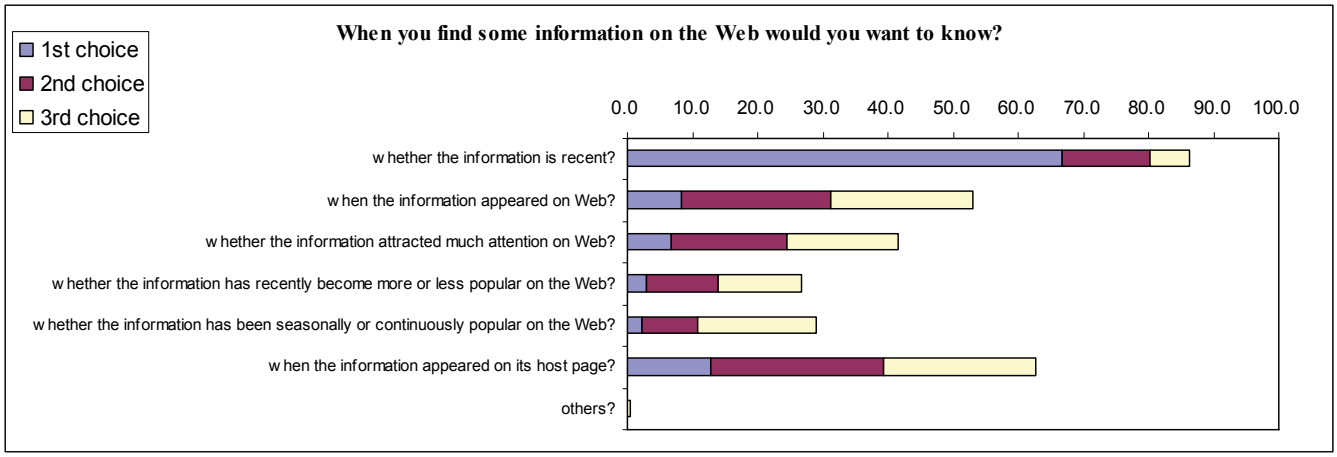

Fig. 7 Questions related to types of temporal support for information found on Web.

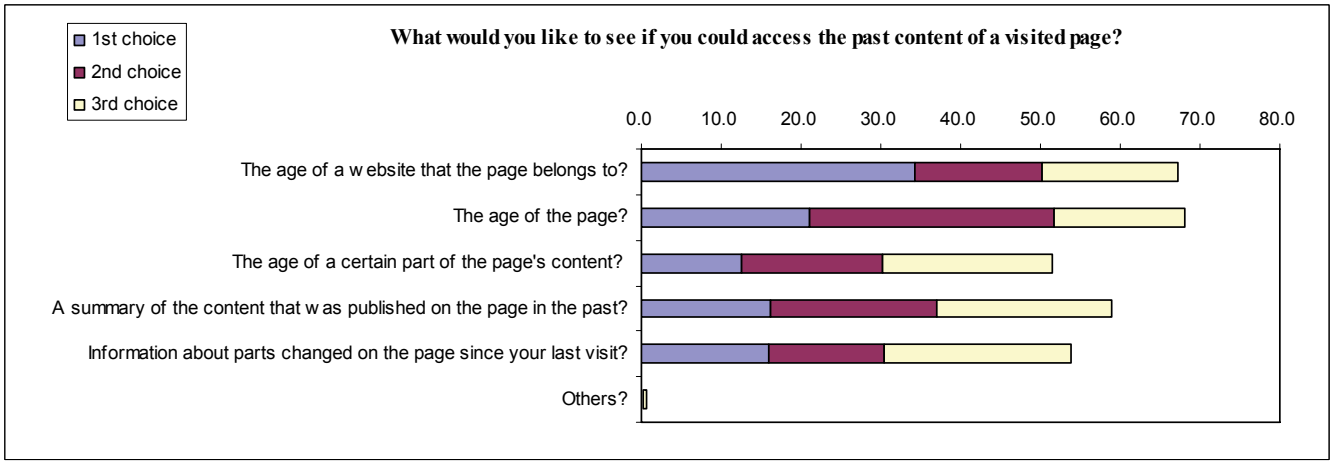

Fig. 8 Questions related to temporal support for visits to pages (specific needs).

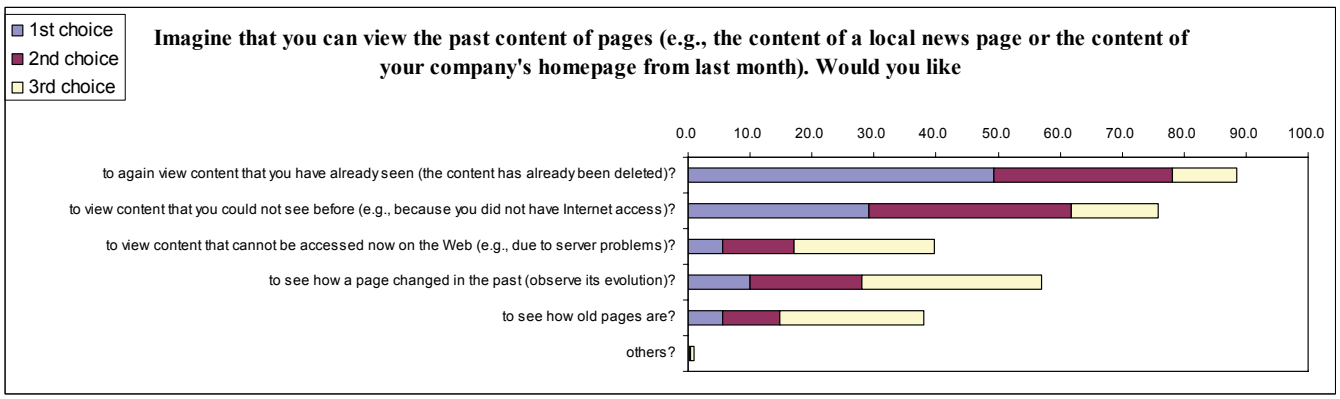

Fig. 9 Questions related to temporal support for visits to pages (specific situations).

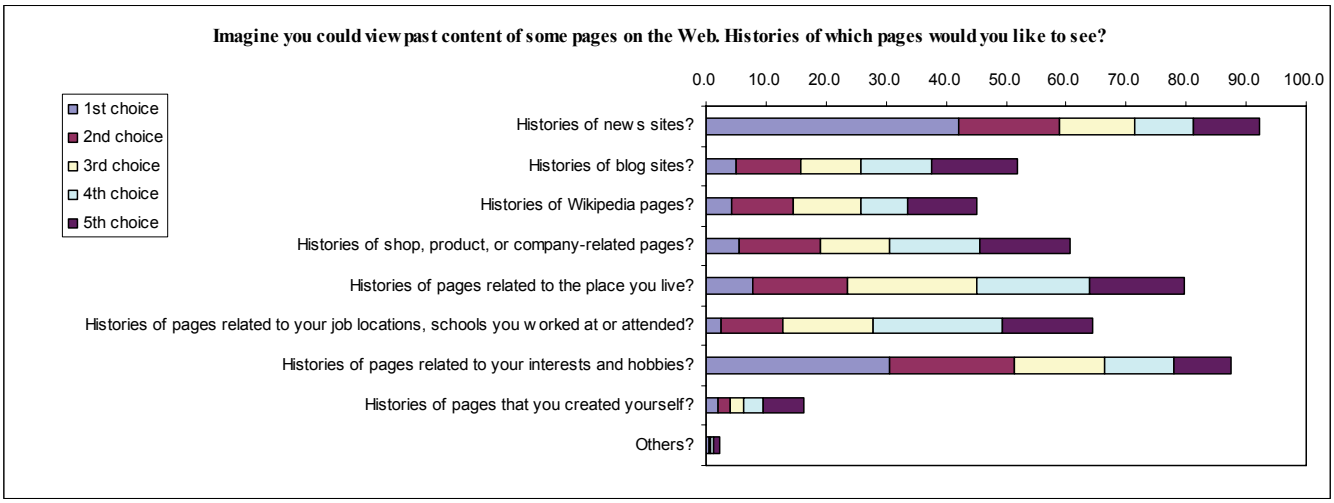

Fig. 10 Questions related to types of pages for which temporal support is required.

We need to emphasize here that the questionnaire we administered suffered from certain weaknesses such as its obvious inability to ask subjects in person and the generalized nature of the questions it asked; however, we do think that it provides some basic ideas on the needs and wishes of average users regarding past Web content.

6. DISCUSSION

6.1.1 User Interaction Issues

The types of pages for which users want to see historical data can vary from person to person. The depth to which users would like to have interaction with page histories also depends on various cases. Intuitively, a rule is likely to apply that the further one goes back in a document's history the less its historical content has a relation to its present content. This can be managed to some extent by providing implicit information to users on the age of visited snapshots or by introducing decaying weights to the process of generating the history summary.

Cost is another important aspect of history-based interaction. Fetching data from Web archives is both time and resource consuming. The cost should naturally decrease as archive infrastructures are restructured and developed. Nevertheless, users should be made aware of this cost, e.g., when determining the length and number of samples required to create historical summaries.

6.1.2 Legal Issues

An important issue is related to the copyrights of content and the attitude of its providers to archiving. Some content owners may not allow providing open access to their past data. The reasons for this are various. As the content is obsolete owners may be concerned that some users might mistake it for current content, hence receiving out-dated or wrong information or advice. The possibility of comparing historical information with actual information on various real-world objects, such as companies, may also be a concern to certain content providers, especially, if they prefer to hide specific information. Simply speaking, users would be entrusted with more power to assess the quality and characteristics of real-world objects, be it companies, institutions, or persons. Another reason is that some portion of the original traffic to sites may be captured by Web archives, thus, limiting the popularity or revenue of the site. Lastly, some Web sites, especially news providers, have introduced systems to charge customers for access to their archives.

The Internet Archive currently enables content owners to forbid content archiving and re-publishing by direct contact or by inserting "no-archive" information to document metadata. Some countries, on the other hand, enforce unrestricted archiving conducted by specialized institutions according to data-deposition laws, yet, at the same time they may limit access to this data only to particular groups of users, such as historians or researchers [14].

We believe that end users should decide what types of data should be preserved and what types of access should be provided to gain entry to such information in order to make it popular and useful. From the viewpoint of the service provider, it is important to deliver explicit information on the age and provenance of accessed or analyzed data. When confronted by intense user interest, various institutions, lawmakers, and other entities would probably make compromises to satisfy Web users.

6.1.3 Archiving Issues

One important issue is related to the process of selecting content for archiving. Since the Web is enormous, can change unpredictably, and archiving institutions have limited resources, then only selected parts of the Web can, naturally, be archived and preserved. The remaining parts, if not archived by their content owners or providers, face extinction. The Internet Archive currently seems to select pages for archiving based on how popular pages are with users, which is measured by retrieving browsing statistics from them. The frequency with which snapshots of popular pages are captured is then determined by the estimated frequencies of change for pages.

A related issue is how to preserve Web content when technology and standards are quickly changing and how long to store the data. Also, the integrity and credibility of archived data should be carefully considered. These problems are currently at the center of attention in the Web-archiving community [14].

Lastly, as we previously mentioned, pages may change their URLs over time despite their content basically remaining the

same. The solution to this problem is, however, not trivial and requires a more thorough investigation.

7. CONCLUSIONS

Web pages are currently viewed usually through their present content only. Thus, users usually access them unaware of the histories that the pages have and ignorant of the fact that these histories can be reconstructed from data available in Web archives or other archival repositories. Thus far, researchers have placed little interest on providing temporal support that users would need when browsing the Web. This paper first proposed a tighter integration of documents with their histories and discussed several possible types of interaction that users might wish to have with the histories of Web pages they visited. The categorization we outlined may not, however, be exhaustive as new potentially useful and interesting interaction types could be discovered. The second contribution made by this paper was that it presented the results of a survey conducted with the objective of understanding user needs and behaviors regarding changes in Web documents. The results generally indicated relatively high levels of interest by users in document histories, although, at the same time they revealed less use of Web archives – the main custodians of our Web heritage.

We believe that the availability of novel history-based interaction models and effective applications for supporting them should help to better determine users' actual needs and this should generally make them more aware of the existence of accumulated historical data and the potential derived from their use.

8. ACKNOWLEDGMENTS

This research was supported by the MEXT Grant-in-Aid for Scientific Research in Priority Areas entitled: Content Fusion and Seamless Search for Information Explosion (#18049041, Representative: Katsumi Tanaka), the Kyoto University Global COE Program: Informatics Education and Research Center for Knowledge-Circulating Society (Representative: Katsumi Tanaka) and by the MEXT Grant-in-Aid for Young Scientists B (#18700111, Representative: Adam Jatowt; #18700110, Representative: Yukiko Kawai).

9. REFERENCES

[1] Broder, A. Z. A taxonomy of web search. *SIGIR Forum*, 36(2), 2002, 3–10.

[2] Brusilovsky, P., Kobsa, A., and Nejdl, W. *The Adaptive Web, Methods and Strategies of Web Personalization*, Berlin Heidelberg New York, Springer, 2007.

[3] Dyreson C. E., Lin H.-L., and Wang Y. Managing versions of web documents in a transaction-time Web server. *Proceedings of the 13th International World Wide Web Conference*, 2004, 422–432.

[4] Francisco-Revilla, L., Shipman, F.M., Furuta, R., Karadkar, U. and Arora, A. Perception of content, structure, and presentation changes in Web-based hypertext, *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia*, 2001, 205–214.

[5] Freyne, J., Farzan, R., Brusilovsky, P., Smyth, B. and Coyle, M. Collecting community wisdom: integrating social search & social navigation. *Proceedings of the International Conference on Intelligent User Interfaces*, 2007, 52–61.

[6] Haake, A. and Hicks, D. VerSE: Towards hypertext versioning styles. *Proceedings of the 10th ACM Conference on Hypertext and Hypermedia*, 1996, 224–234.

[7] IIPC's Access Working Group. *Use cases for Access to Internet Archives*, 2006, http://netpreserve.org/publications/iipc-r-003.pdf

[8] Jatowt, A., and Ishizuka, M. Web page summarization using dynamic content, *Proceedings of the 13th International World Wide Web Conference*, 2004, 344–345.

[9] Jatowt, A., Kawai, Y., Nakamura, S., Kidawara, Y. and Tanaka, K. Journey to the past: proposal for a past web browser. *Proceedings of the 17th ACM Conference on Hypertext and Hypermedia*, 2006, 134–144.

[10] Jatowt, A., Kawai, Y. and Tanaka, K. Detecting age of page content. *Proceedings of the 8th International Workshop on Web Information and Data Management*, 2007, 137–144.

[11] Jatowt, A., Kawai, Y. and Tanaka, K. Visualizing historical content of web pages, *Proceedings of the International World Wide Web Conference*, poster, 2008.

[12] Jatowt, A., Kawai, Y. and Tanaka, K. Personalized detection of fresh content and temporal annotation for improved page revisiting, *Proceedings of the 17th Conference on Database and Expert Systems Applications*, 2006, 832–841.

[13] Luesebrink, M.C. The moment in hypertext: a brief lexicon of time. *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia*, 1998, 106–112.

[14] Masanes, J. (ed.). *Web archiving*. Berlin Heidelberg New York, Springer Verlag, 2006.

[15] McCown, F., Diawara, N. and Nelson, M.L. Factors affecting website reconstruction from the web infrastructure. *Proceedings of the Joint Conference on Digital Libraries*, 2007, 39–48.

[16] Nelson, T. H. *Literary Machines*, edition 87.1, Sausalito, CA, USA, Sausalito Press, 1987.

[17] Obendorf, H., Weinreich, H., Herder, E., and Mayer M. Web page revisitation revisited: implications of a long-term click-stream study of browser usage. *Proceedings of Conference on Human Factors in Computing Systems,* 2007, 597-606.

[18] Shipman F. M. and Hsieh H. Navigable history: a reader's view of writer's time. *New review of hypermedia and multimedia*, vol. 6, 2000, 147–167.

[19] Wexelblat, A, and Maes, P. Footprints: history-rich tools for information foraging. *Proceedings of Conference on Human Factors in Computing Systems*, 1999, 270–277.

[20] Vitali, F. Versioning hypemedia. *ACM Computing Surveys* 31(4): 24. ACM Press, 1999.

[21] Yanbe, Y., Jatowt, A., Nakamura, S., and Tanaka, K. Can social bookmarking enhance search in the web? *Proceedings of the Joint Conference on Digital Libraries*, 2007, 107–116.