Method of Retrieving a Web Browsing Experience Using Semantic Periods

Tetsushi Morita[1], Tsuneko Kura[2], Tetsuo Hidaka[1],
Akimichi Tanaka[1], and Yasuhisa Kato[1]

[1] NTT Cyber Solutions Laboratories, NTT Corporation
[2] NTT Service Integration Laboratories, NTT Corporation,
3-9-11 Midori-cho Musashino-shi, Tokyo, Japan 180-8585
{morita.t, kura.tsuneko, hidaka.tetsuo, tanaka.akimichi, kato.yasuhisa}@lab.ntt.co.jp

Abstract We describe a system which reminds users of not only the content of web pages but also which web pages we viewed in a session, why the pages were viewed, and what knowledge we acquire from them. We developed a module for automatically collecting user logs. We focus on meaningful period, which metadata is added to by extracting the log data. We call this type of period a "semantic period". One type of semantic period is an "intensive period" which includes a lot of activity related to keywords, and the other type of period is an "active period" when a web page is actively shown in a window. We also describe an interface for retrieving information acquired in a browsing experience using the semantic period. Finally, we show that our method lets the user recall important knowledge based on the web pages viewed in the past more efficiently.

Keywords: semantic desktop, information retrieval, web browsing history, user interaction

1 Introduction

Most of us have retrieved the same web page more than once after remembering that the experience was useful. A report says that the rate of previously seen web pages among all pages that people viewed was 81% [2]. However, we sometimes start searching for those web pages from scratch or cannot rediscovery them using an Internet search engine, even though we viewed them in the past.

A desktop search system enables us to retrieve a previously seen web page by keyword matching [3][4]. However, the information that we acquire in an experience, such as web browsing, is not limited to the content of web pages. We seem to recognize which web pages we viewed in a session, why the pages were viewed, and what knowledge we obtained from them. Such a variety of information is difficult to acquire if we only retrieve the content of web pages.

We think that user operation history related to a web page is personal metadata about the page. Our proposed method collects user operation logs which contain basic data such as events of activating window, copying, and printing. We focus on meaningful period, which metadata is added to by extracting the log data. We call this type of period a "semantic period". One type of semantic period is an "intensive

period," which includes a lot of activity related to keywords, and the other type of semantic period is an "active period" when a web page is actively shown in a window. Details of these periods are described later. Using the semantic period and the web page text, our method reminds us of the various kinds of information that we acquired simultaneously in the past.

2 Related work

A lot of research studies collect log data from personal computers (PC) and make efficient use of it.

Several semantic desktop search systems have been proposed. One of the systems helps a user to retrieve web pages and e-mails according to context information by adding metadata such as an url of web page visited subsequently and the destination address of an e-mail [1]. MyLifeBits proposed some methods to add keywords easily as metadata to files in PCs [5]. A previous version of our method helps users to retrieve web pages viewed in the past by calculating the personal importance of web pages by using log data from PCs [9]. These desktop search systems and those described in the foregoing chapter mainly aim to find web page content viewed in the past efficiently. The aim of previous methods is different from that of the proposed method, which aims to find semantic periods.

Methods to remind user of acquired information by finding timing of events such as viewing a web page or creating a file have been studied. The history function of popular web browsers and time-view function of desktop search systems show web pages in chronological order of download time. The Smart Back system supports an advanced backtracking function that extracts important web pages using the browsing sequence, retrievals, and bookmark settings for example [6]. TimeMachineComputing saves the condition of an application by simulating a desktop screen and allowing a user to view past conditions of the application. Namely, the user sees icons of files and computer post-it messages on the desktop. [7]. Memory Organizer visualizes the relationships of collocations among words of anchor text clicked within a period specified by the user [8]. These methods remind users of information acquired in the past according to the specified time. For retrieving the times of user actions, these methods provide several functions like input of keywords and filenames, and a visual interface of the browsing history. However, our method is different from those because it enables a user to retrieve time intervals instead of only times.

3 Information acquired through web browsing experience

One type of information acquired through a web browsing experience is the knowledge that a certain page exists and its address, as typified by the URL, that enables it to be found again. People often visit not only one web page but multiple web pages when searching on a theme. For example, a user might visit the web page of a product after he read a reliable page recommending the product. The relationship among the web pages he visited is also information acquired through the web

browsing experience. What knowledge he got from a web page is also the information acquired through the web browsing experience. Namely, he acquired various kinds of information through the web browsing experience. We chose to target the following information, called "obtained information", among the various types of information acquired through the web browsing experience. Obtained information is:

- Existence and address of web pages
- Relationships among the web pages
- Knowledge obtained from the web pages.

Our method aims to remind the user of the obtained information efficiently. Namely, it should:

- remind him about a lot of obtained information
- remind him about it quickly
- remind him about important obtained information

4 Proposed method aiming at semantic periods

Many people retrieve information using keywords. If they retrieve a web page using keywords, they often learn of a web page's existence and its address efficiently. But, it seems to be difficult to remind users efficiently about a lot of obtained information because they need to choose and visit many retrieved independent web pages. So, we focus on meaningful period called "semantic period" which metadata is added to. We propose a method of retrieving a period, which includes a lot of activity related keywords. We call this period the "intensive period". We assume that a lot of obtained information is also contained in an intensive period. For example, if a user has a web browsing experience to research a product, he finds only an intensive period related to this experience and then he acquires a lot of obtained information such as the existence and address of multiple web pages that were visited at that time. By chronologically tracing his activities, such as which web pages he viewed in the intensive period, he acquires the relationships among the pages and the knowledge that he got from the pages. To remind him of the information he obtained in an intensive period efficiently, we target another period called the "active period" and we propose a method that presents a user's activities using the active periods in an intensive period.

To provide the method using semantic period, the system must collect the history of actions on the user's PC, analyze the history to extract the semantic period with added metadata, and present it with effective interaction (Fig. 1). In this section, we describe a module for automatically collecting a detailed history of user actions, methods for extracting the active and intensive periods, and interfaces for reminding the user of obtained information using the semantic period.

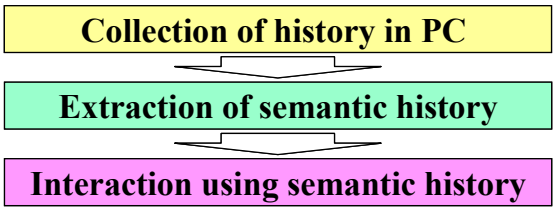

Fig. 1. Steps for using semantic history.

4.1 Automatic logging module

It is difficult to force a user to perform the actions required to create history data such as when and how he or she viewed a web page. So we developed a logging module that automatically saves detailed logs in the user's PC (Fig. 2). The automatic logging module monitors the event messages of an operation system (OS), so it does not depend on applications. Specifically, it collects the PC's mouse, keyboard, copying, and printing event and window conditions. But detailed log data that depends on an application such as URL and source files cannot be saved in the form of operating system events. So we developed an additional module for Microsoft Internet Explorer (IE) that monitors IE's event messages and properties. Specifically, this additional module collects URLs that IE shows, source files, thumbnails, http headers, and texts selected by the user. The logging module has an encryption function to protect the user's privacy.

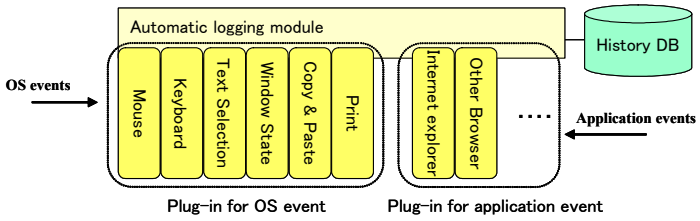

Fig. 2. Automatic logging module.

4.2 Extraction for semantic period

4.2.1 Active period

The logging module saves a huge amount of data. To ensure that the data is useful, we focus on the active period and extract meaning from the data. This section explains the active period, the degree of user's attention to the active period, and the degree of importance of the active period to keywords. Most PCs these days use a graphical window-based OS and web pages are displayed in a browser window. We can change

web pages by clicking a link in the active window or by changing windows or tabs within the same window. For example, we sometimes search for interesting web pages by clicking a link. Then, if we find some interesting web pages, we open multiple windows of IE and shift the windows so that we can compare web pages. Therefore, we define the active period as the period when a web page is actively shown in a window. What it comes down to is that if the viewed web page is changed by clicking a link or shifting the active window (tab), a new active period begins. We regard the active period as a useful unit in which we integrate log data. As the attributes of an active period, we define the starting and ending times of the activity, a hash of the source file, window information, URL, page title, referrer, anchor text clicked on a last web page, related search query, path of a thumbnail, path of the source file, encoding of the source file, and actions such as keyboard inputs, mouse clicks, and printing events.

In an active period, we perform various actions such as viewing a web page, copying interesting sentences, and printing the web page if we judge it to be especially interesting. We estimate the degree of user's attention to the active period on the hypothesis that if a user performs many actions in the active period, then he/she is paying a lot of attention in that period. Specifically, the degree of user's attention to the active period is calculated by eq. (1). Moreover, we estimate the degree of importance of the active period to a set of keywords because a user seems to retrieve the information in his/her experience related to keywords. Specifically, the degree of importance of the active period to a set of keywords is calculated by eq. (2). These degrees are also regarded as properties of the active period.

$$Att(ap) = \sum_i (E_i \times Fr_i) \qquad (1)$$

$$IA(k, ap) = Att(ap) \times R(k, ap) \qquad (2)$$

$Att(ap)$: Degree of user's attention to the active period.
$IA(k, ap)$: Degree of importance of the active period to the set of keywords.
$R(k, ap)$: Relevance ratio of a web page in an active period to set of keywords k
E_i : Weighting factor of action category i.
Fr_i : Number of occurrences of action category i in an active period ap
ap : Active period
k : Set of keywords
i : Action category

The relevance ratio of a web page in an active period to the set of keywords is given the value of TF-IDF based on the set of all web pages logged by the automatic logging module. For action categories, we provide the amount of active time, copying, printing, mouse clicking, keyboard input, and text selection.

4.2.2 Intensive period

This section explains the intensive period, which includes a lot of activity related to a set of keywords, and the degree of importance of the intensive period to a set of keywords. If a user views web pages related to keywords in a concentrated manner, most but not all of them include the keywords. There are some cases where a few web pages related to the keywords do not actually include the keywords. The user occasionally views web pages that deviate from the keywords for a short time and

then returns to viewing web pages that are related to the keywords. Therefore, we extracted the intensive period through the following steps.

First, the degree of importance of the random time t to the set of keywords k is calculated by eq. (3). This means that the degree is calculated by dividing the degree of importance of the active period, including the time t, to the set of keywords by the length of the active period. Then, the average degree of importance of the random time t to the set of keywords k is determined from eq. (4). Here, α is a parameter for averaging. If the average degree is not more than parameter β, the operation related to the keywords is regarded as being discontinuous at time t (Fig. 3). Namely, a period that has a continuous value of 1 for the value of function $B(k,t)$, which judges continuance, is regarded as an intensive period (eq. (5)). The values of α and β are decided by heuristics.

In this way, the method can regard a period as an intensive period using activities before and after in time, even if the user viewed some web pages that did not include the keywords for a short time during the period.

$$IT(k,t) = IA(k,ap)/(apet - apst) \qquad (3)$$

$$AIT(k,t) = \int_{t-\alpha}^{t+\alpha} IT(k,t)/2\alpha \; dt \qquad (4)$$

$$B(k,t) = \begin{cases} 0, & if \quad AIT(k,t) \le \beta \\ 1, & otherwise, \end{cases} \qquad (5)$$

$apst$: Start time of the active period ap

$apet$: End time of the active period ap

$IT(k,t)$: Degree of importance of time t to the set of keywords k

$AIT(k,t)$: Average degree of importance of time t to the set of keywords k

Next, the degree of importance of each extracted intensive period ip to the set of keywords k is determined by eq. (6). An intensive period whose degree of importance is high includes a lot of activities related to the set of keywords k, so presenting activities in this period can remind the user of a lot of obtained information. As the attributes of an intensive period, we define the starting and ending times of a period and degree of importance of intensive period ip to the set of keywords k.

$$II(k,ip) = \int_{st}^{et} IT(k,t)dt \qquad (6)$$

$II(k,ip)$: Degree of importance of intensive period ip to the set of keywords k

st : Start time of the intensive period ip

et : End time of the intensive period ip

ip : Intensive period

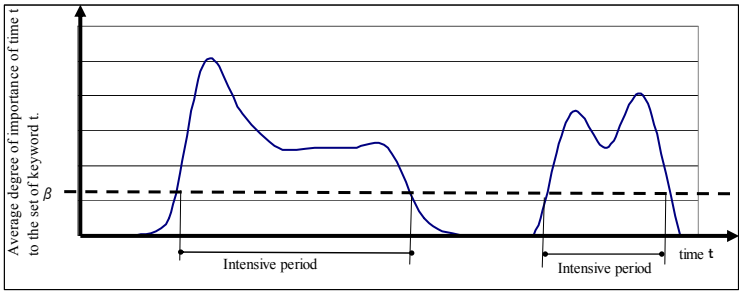

Fig. 3. Extraction of intensive periods.

4.3 Interaction

Using the semantic periods, the system provides interactions to remind a user of his/her obtained information. This section describes interfaces for retrieving intensive periods related to keywords and for understanding user activities in a periods. In the normal situation, we assume that the user first retrieves an intensive period and then analyzes his/her activities during that period.

4.3.1 Interface for retrieving intensive periods

The interface for retrieving intensive periods related to keywords is shown in Figure 4. If the user inputs a set of keywords, the intensive periods are shown in order of importance to the set of keywords or in chronological order. The characteristics of each intensive period are shown, so the user can recall an overview of his/her activities during the intensive period. Specifically, the characteristics are the start and end time of the intensive period, the degree of importance of the intensive period to the set of keywords, keywords input into an Internet search page, words that occur frequently on web pages viewed in the intensive period, and thumbnails with titles of the five most important web pages. The importance of a web page is calculated by summing the degrees of importance of active periods to the set of keywords whose URL attributes are the same in the intensive period.

If the user uses this interface, he/she can easily find an intensive period that is important to a set of keywords. Then the user is reminded of the obtained information via another interface described in next section.

Fig. 4. Interface for retrieving intensive periods.

4.3.2 Interface for understanding activities in an intensive period

The user interface for understanding activities in an intensive period is shown in Figure 5. The sequence of active periods is displayed in chronological order. Each active period is represented by its thumbnail attribute. Therefore, a user can remember a number of web pages by following the sequence.

In the interface, anchor text clicked on the previous web page is shown above the active period. In web browsing, we often input a query at an Internet search engine site to find interesting web pages. If its URL attribute is the results page of the search engine, then the query input to the search engine is used instead of the anchor text. Therefore, users can not only understand the content of a web page more clearly but also be reminded why they viewed it. For example, they might be reminded of the relationships among the web pages that they used to start a survey of thin digital cameras by inputting “digital camera thin type” into an Internet search engine, get recommendations by reading special topics on a news site, and compare detailed specifications of digital cameras on official sites. To help users remember for a short time the large amount of important information acquired in such an experience, the system marks active periods whose degrees of importance to keywords input by the user are higher than those of other active periods by surrounding them with red or orange boxes. Conversely, active periods whose degree of importance to keywords is low can be filtered out because we assume that they were not useful. This interface is useful to understand activities in not only an intensive period but an ordinary period.

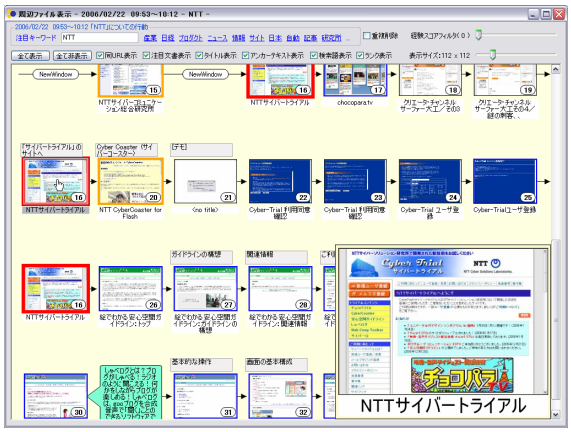

Fig. 5. Interface for understanding activities in an intensive period.

In the real world, people sometimes underline or highlight important sentences when they read a book. They can also easily select important sentences using a mouse on a web page shown in a web browser. The green callout from a thumbnail in Fig. 5 show sentences that the user highlighted in the past. If the user clicks a thumbnail with a callout, the web page with the selected sentences highlighted is displayed (Fig. 6). By viewing the highlighted web pages, the user can easily find important sentences and be reminded of the knowledge obtained from those pages.

Fig. 6. The system highlights sentences that the user previously selected.

5 Evaluation

We evaluated whether our method efficiently reminds users of obtained information. In this evaluation, we selected only "knowledge obtained from web pages" as obtained information because it was the final result of the user's web browsing experience. For example, when a user researched about Ishigaki island (in Japan) on the web, this evaluation experiment checked whether the system reminded the user not of the URL "www.ishigaki.tmp.com/tmp.html" but of the knowledge that "It is difficult to go surfing on Ishigaki island". The evaluation considered whether the reminded knowledge was important or not.

5.1 Conditions

(1) Comparison method
In this experiment, the proposed method was compared with a conventional method that shows web pages viewed in chronological order [3]. Specifically, the "timeline view" of this conventional method shows web page titles in chronological order. The timeline view shows the attributes of a web page such as download date, title, and URL. It also enables users to retrieve viewed web pages by set of keywords. The results of the retrieval are shown in chronological order or in order of relevance to the keywords. This view shows the attributes of a web page such as a thumbnail, a snippet, download time, a title, and an URL.

(2) Test subjects
In this experiment, the subjects were not novice computer users because they needed to search for information on the web. Details about the subjects are shown in Table 1.

Table 1. Details about subjects.

No. of subjects	20
Ages	20s to 40s
Sex ratio	1:1
Frequency of PC use	1–7 days per week
Familiarity with each method	Subjects practiced each method for half an hour

(3) Parameters of active period
We chose E_i, which is a weighting factor of the action category, to calculate the degree of user's attention to the active period by heuristics (Table 2).

Table 2. Value of weighting factor of action category.

Action category (unit)	Weighting factor
Viewing web page (s)	0.17
Copying (times)	10
Selecting (times)	50
Printing (times)	50
Mouse (times)	0.1
Keyboard (times)	0.1

5.2 Procedure

The experimental procedure can be divided into two steps. The first step was the experiencing step in which subjects obtained information from new web pages. The second step was the reminding step in which they were reminded of the obtained information by either our new method or the conventional method (Fig. 7). We analyzed the experimental results and compared the effectiveness of our method with that of the conventional method.

(1) Experiencing step
The subjects browsed web pages freely for half an hour to solve a given task and reported the knowledge that they obtained from the web pages. The ten most important items of knowledge were marked. Activities during the browsing were logged by the automatic logging module. The given tasks were "survey retrieval task [10]" designed by us. Examples of the task and the knowledge reported by a subject are shown in Table 3.

(2) Reminding step
One week after the experiencing step, subjects were reminded of the knowledge they obtained from web pages by using the interface for understanding activities in a period (our new method) or the conventional method. The subjects wrote out the reminded knowledge on their answer-sheet as soon as they reacquired it during a period of fifteen minutes. The knowledge reminding time included this writing time. The knowledge could be distinguished by how quickly they were reminded of the previously obtained knowledge. We categorized the time taken as: 0–4, 4–7, 7–10, or 10–15 min.
To avoid subject/task bias in the results, we chose to have subjects perform tasks with both methods. Namely, each of the subjects answered about his/her knowledge for 4 tasks and each subject used our method for 2 tasks and the conventional method for 2 tasks. Tasks were assigned randomly. In total, we received 40 written answer-sheets for each method.

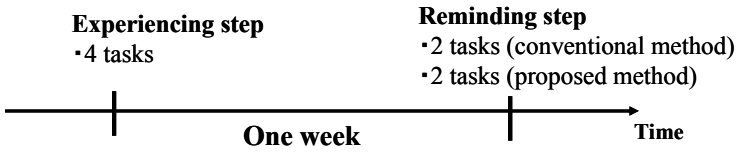

Fig. 7. Procedure of the experiment.

Table 3. Examples of given tasks and reported knowledge.

Example of given tasks
You decide to have a vacation on an isolated island for three days. Please survey information related to the following islands: Tanega Island, Ishigaki Island, and Sado Island (in Japanese).
Examples of leaved knowledge
It is difficult to go surfing on Ishigaki Island.
It costs 9410 yen to get from Kagoshima to Tanega Island.
The beach at Komaza is attractive.

5.3 Results

To evaluate how efficiently each method reminded subjects of the knowledge they obtained from web pages, we calculated the following score (eq. (7)). The knowledge marked as important knowledge was given twice the weight of non-marked knowledge. Reminded knowledge that had not been reported in the experiencing step was excluded.

$$Score = (2a + b)/(2c + d) \qquad (7)$$

a: Number of reminded items of important knowledge
b: Number of reminded items of ordinary knowledge
c: Number of reported items of important obtained knowledge
d: Number of reported items of ordinary obtained knowledge

The scores for the proposed and conventional methods are shown in Fig. 8. Our method reminded subjects of "knowledge obtained from web pages" more efficiently than the conventional method. Using an unpaired t-test whose significant level was 5%, the t-test values for 4, 7, 10, and 15 minutes were 3.88, 4.97, 5.73, and 4.04, which mean that the proposed method was statistically significant at reminding

subjects of the knowledge they obtained from web pages compared with the conventional method.

The scores per minute in each period (0–4, 4–7, 7-10, and 10–15 minutes) are shown in Fig. 9. For 0–10 minutes, the values for our method are especially higher than those for the conventional method, which shows that our method reminded subjects of the knowledge they obtained from web pages at the beginning of the reminding period. In this experiment, users were reminded of knowledge that had been obtained during the thirty-minute experiencing step. We assumed that users do not usually take the same amount of time (thirty minutes) to be reminded of this knowledge in the reminding step. Users want to be reminded quickly, and our method meets this demand more efficiently.

We investigated why our method reminded users of knowledge more efficiently. One of the reasons is that users are reminded of knowledge contained in the multiple web pages viewed in active periods, which means that the users took notice of them. Another reason is that it enables users to remember what they were thinking during an experience because users understand the attributes of the active period such as its thumbnail, related search query, and anchor text clicked on a last web page. A third reason is that it takes a short time to remind users of knowledge after they find that it exists and find the address of a web page containing it, because selected sentences are clear. The sequence of active periods reminds users in more detail about their thinking during an experience than the sequence of downloaded web pages in chronological order.

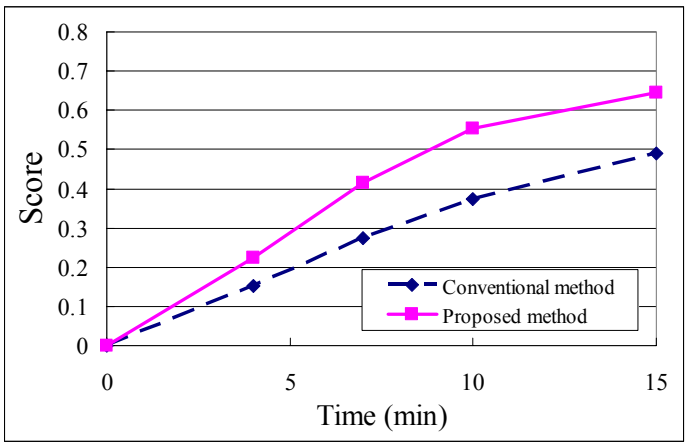

Fig. 8. Score of conventional and proposed methods.

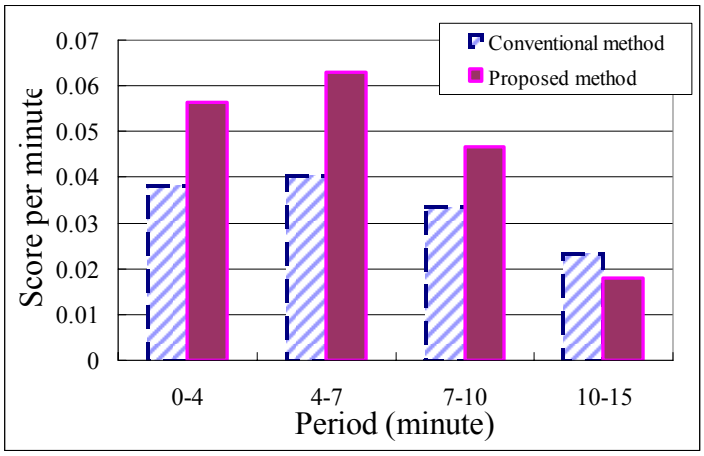

Fig. 9. Score per minute.

6 Conclusion and future work

We proposed a method of efficiently reminding a user of information obtained through a web browsing experience by retrieving that experience. This method retrieves not the viewed web pages but also the semantic periods by installing an automatic logging module in the user's PC that collects fundamental history data. We extracted semantic history data as an active period with its attributes when a web page was actively viewed in a window and as an intensive period that includes a lot of activity related to keywords. We extracted the attributes of the intensive period as the degree of importance of the intensive period to keywords. We designed interfaces to retrieve intensive periods and to understand activities in the intensive period using active periods with their attributes. Our method makes a user efficiently remember obtained information which is the existence and address of web pages, relationships among the web pages, and knowledge obtained from the web pages. An evaluation experiment showed it reminded subjects of knowledge obtained from web pages more efficiently than the conventional method that shows viewed web pages in chronological order.

In the future, we plan to open up access to the application interface for extracting active periods, intensive periods, and their attributes to other desktop application developers. We will evaluate this method of extracting intensive periods and the interface for retrieving intensive periods. We will consider how to decide parameters for extracting the intensive period and calculating the degree of user's attention to the active period. For example, we can apply a method of machine learning to enable the system to adapt to the user's situation and develop an user interface for adjusting the parameters. The automatic logging module has already collected history related not only to web pages but also e-mails and image files. We plan to support active periods

of not only web pages but also e-mails, office documents, and other files. A method of summarizing a sequence of active periods in an intensive period will be studied.

References

1. Paul Chirita, Rita Ghita Alexandru Gavriloaie, Stefania, Wolfgang Nejdl, Paiu Raluca: Activity Based Metadata for Semantic Desktop Search. 2nd European Semantic Web Conference (ESWC05) (2005).
2. A. Cockburn and B. McKenzie: What Do Web Users Do? An Empirical Analysis of Web Use. Int. J. Human-Computer Studies, 54(6), (2001) 903-922.
3. About Google Desktop, http://desktop.google.com/about.html
4. Yahoo! Desktop Search, http://desktop.yahoo.com/
5. Gim Gemmell, Gordon Bell, Roger Lueder, Steven Drucker, and Curtis Wong: Mylifebits: Fulfilling the memex vision. In ACM Multimedia'02 Proceedings, (2002) 235–238.
6. Natasa Milic-Frayling, Rachel Jones, Kerry Rodden, Gavin Smyth, Alan Blackwell, Ralph Sommerer: SmartBack: Supporting Users in Back Navigation. 13th International World Wide Web Conference (WWW2004), New York City NY (2004).
7. Jun Rekimoto: Time-Machine Computing: A Time-centric Approach for the Information Environment. ACM UIST'99 (1999).
8. Harumi Murakami and Takashi Hirata: A System for Generating User's Chronological Interest Space from Web Browsing History. International Journal of Knowledge-Based and Intelligent Engineering Systems, Vol. 8, No. 3, (2004) 149-160.
9. T. Morita, T. Hidaka, T. Kura, K. Ooura, Y. Kato: Desktop search system based on the Action-Oriented algorism. Proc. APSITT2005 (6thAsia-Pacific Symposium on Information and Telecommunication Technologies), (2005) 204-207.
10. Koji Eguchi, Keizo Oyama, Emi Ishida, Noriko Kando, and Kazuko Kuriyama: Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure. IEICE Transactions on Information and Systems, Vol. E86-D, No. 9, (2003) 1804-1813.