System for Reminding a User of Information Obtained through a Web Browsing Experience

Tetsushi Morita
NTT Corporation
NTT Cyber Solutions Laboratories
3-9-11 Midori-Cho
Musashino-shi Tokyo,
180-8585 Japan
81-422-59-4840
morita.t@lab.ntt.co.jp

Tetsuo Hidaka
NTT Corporation
NTT Cyber Solutions Laboratories
3-9-11 Midori-Cho
Musashino-shi Tokyo,
180-8585 Japan
81-422-59-7150
hidaka.tetsuo@lab.ntt.co.jp

Akimichi Tanaka
NTT Corporation
NTT Cyber Solutions Laboratories
3-9-11 Midori-Cho
Musashino-shi Tokyo,
180-8585 Japan
81-422-59-4483
tanaka.akimichi@lab.ntt.co.jp

Yasuhisa Kato
NTT Corporation
NTT Cyber Solutions Laboratories
3-9-11 Midori-Cho
Musashino-shi Tokyo,
180-8585 Japan
81-422-59-4420
kato.yasuhisa@lab.ntt.co.jp

ABSTRACT

We propose a system for reminding a user of information obtained through a web browsing experience. The system extracts keywords from the content of the web page currently being viewed and retrieves the context of past web browsing related to the keywords. We define the context as a sequence of web browsing when many web pages related to the keyword were viewed intensively because we assume that a lot of information connected to the current content was obtained in the sequence. The information is not only what pages you viewed but also how you found those pages and what knowledge you acquired from them. Specifically, when you browse web pages, this system automatically displays a list of the contexts judged to be important in relation to the current web page. If you select the context, details of the context are shown graphically with marks indicating characteristic activities.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval model, Search process*

General Terms: Algorithms, Management, Design.

Keywords: Context, Information Retrieval, User's Behavior, History.

1. INTRODUCTION

Have you ever been frustrated at failing to rediscover useful web pages that you viewed in the past? We forget and waste various information that we obtain through our own web browsing. Most of us have retrieved the same web page more than once. One report says that the retrieval rate of previously seen web pages among all pages that people view is 81% [3]. The information that we obtain through 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, how the pages were found, and what knowledge we acquired from them. We define the information as "obtained information". In this paper, we describe a system that aims to remind a user of previously obtained information efficiently. The bookmark function of a web browser and a desktop search system are popular to help us to retrieve previously seen web pages by keyword matching [2]. Several semantic desktop search systems have been proposed. One of them helps a user to retrieve web pages and e-mails according to intimate information by adding metadata such as the URL of a web page that is visited subsequently and the destination address of an e-mail [1]. A previous version of our method helps users to retrieve web pages viewed in the past by calculating their personal importance by using log data from personal computers [4]. These desktop search systems help a user to find a web page viewed in the past efficiently. However, they do not seem to be so good at reminding us of the various kinds of information that we acquired simultaneously in the past web browsing experience because we need to choose and visit many retrieved independent web pages.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

2. PROPOSED SYSTEM

We focus on the context in the past. The context is a sequence of web browsing when many web pages related to the content of the web page currently being viewed were viewed intensively and a lot of actions were performed. We call the time of this sequence an "intensive period". We assume that a lot of obtained information is also contained in the context. For example, if a user is researching a product, he first finds the context related to current web pages and then he acquires a lot of obtained information such as the URLs of multiple web pages that were visited at that time. By chronologically tracing his activities, such as which web pages he viewed in the context, he learns how to find the pages and what knowledge he got from them.

2.1 Collecting action logs

It is difficult to force a user to perform the actions required to create history data such as recording when and how he or she viewed a web page. A logging module collects the information about a computer's mouse, keyboard, copying, and printing events and window conditions, the URLs of visited pages, source files, thumbnails, http headers, text selected by the user, and so on. It has an encryption function to protect the user's privacy.

2.2 Extracting keywords of current web page

Our system analyzes the content of the current web page to extract keywords that represent the web page. Its technique is very simple. An analyzer of a browser component obtains the current content C and characterizes it by extracting the most frequent terms. A score S_t is then determined for each term $t \in C$,

where $S_i = (1 + \alpha_0 + \ldots \alpha_n) R(t_i)$ and α_i is a weighting coefficient that varies heuristically with the locality of t_i. That is, extracted terms instanced in anchor text are assigned a higher weight than those in <*body*> text, but a lower one than those in <*title*> or <*h1*>. This $R(t_i)$ is a variation of the *TF-IDF* algorithm, where the term frequency of t_i is multiplied by the inverse document frequency of t_i to approximate each term's importance. Next, the top N ranked characterization terms with weights where stemming and other adjustments are applied are regarded as a set of representative keywords of the current web page. We call this set of keywords the "current keywords".

2.3 Extracting past contexts

We extract an intensive period through the following steps. First, the degree of importance I of a random time t to current keywords k is calculated by eq. (1). We focus on an active period ap when a web page is actively shown in a window. E_i is the weighting factor of action category i such as the amount of active time, copying, printing, mouse clicking, keyboard input, and text selection. Fr_i is the number of occurrences of i in ap. R is a relevance ratio of a web page in ap to k, which is given the value of TF-IDF based on the set of all web pages logged by the logging module. Then, the average degree of importance AI of t to k is determined from eq. (2). Here, α is a parameter for averaging. If the average degree is not more than parameter β, the operation related to the current keywords is regarded as being discontinuous at time t (Fig. 1). The values of α and β are decided by heuristics.

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

$$I(k,t) = \sum_i (E_i \times Fr_i) \times R(k, ap) / (apet - apst) \quad (1)$$

$$AI(k,t) = \int_{t-\alpha}^{t+\alpha} I(k,t) / 2\alpha \, dt \quad (2)$$

apst: start time of the active period
apet: end time of the active period

Next, the degree of importance of each extracted intensive period to current keywords k is determined by eq. (3).

$$II(k, ip) = \int_{st}^{et} AI(k,t) dt \quad (3)$$

st: start time of the intensive period
et: end time of the intensive period

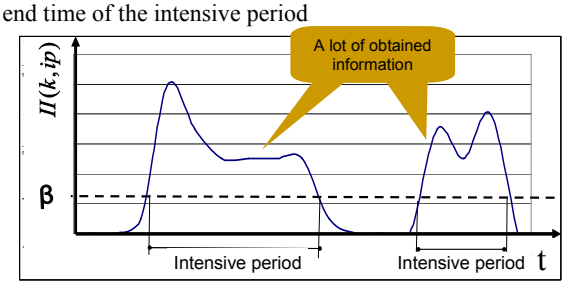

Fig. 1. Extraction of intensive periods.

2.4 Showing details in a context

The user interface for understanding activities in a context is shown in Figure 2. The sequence of viewed web pages is displayed in chronological order by their thumbnails. Therefore, a user can remember several web pages by following the sequence. In the interface, various information such as a query input at an Internet search engine site (Fig. 2(a)), anchor text clicked on the previous web page (Fig. 2(b)), and a title (Fig. 2(c)) are shown. Therefore, users can not only understand the content of a web page more clearly but also be reminded of how 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, getting recommendations by reading special topics on a news site, and comparing detailed specifications of digital cameras on official sites. Selected sentences are highlighted, so the user can be easily reminded of what knowledge was acquired (Fig. 2(d) and Fig. 3). To help users remember for a short time the large amount of important obtained information, the system marks some web pages that are regarded as important ones by surrounding them with red or orange boxes (Fig. 2(e)). Conversely, web pages that are regarded as being unimportant pages can be filtered out (Fig. 2(f)) because we assume that they were not useful.

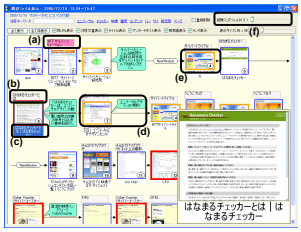

Fig. 2. Details of a context.

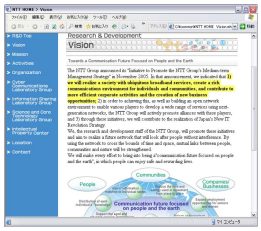

Fig. 3. Highlighted sentences.

3. CONCLUSION

We proposed a system that extracts current keywords from the content of a web page being viewed at present, extracts the context in past web browsing related to the current keywords, and shows details of the context. This system enables a user to be efficiently reminded of various information obtained through his/her web browsing in the past. Specifically, we proposed the context, which is a sequence of web browsing when many web pages related to the current content with a lot of activities were viewed intensively. We described how to extract the context and how to calculate the value of the context as the degree of importance to current keywords. We designed interfaces to understand activities in the context.

4. 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] About Google Desktop, http://desktop.google.com/about.html

[3] 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.

[4] T. Morita, T. Hidaka, T. Kura, K. Oura, Y. Kato: Desktop search system based on the Action-Oriented algorism. Proc. APSITT2005 (6thAsia-Pacific Symposium on Information and Telecommunication Technologies), (2005) 204-207