A Filtering System of Web History Using the Browsing Characteristic

Keita Arai[1(✉)], Makoto Oka[2], and Hirohiko Mori[2]

[1] System Information Engineering, Tokyo City University, Tokyo, Japan
gl48l802@tcu.ac.jp
[2] Department of Industrial and Management Systems Engineering, Tokyo City University, Tokyo, Japan
{moka,hmori}@tcu.ac.jp

Abstract. Most Web browsers have a Web history function to support to go back to the pages which user use to watch. However, only few people use it to revisit Web pages. In this study, we suggest a filtering system of Web history that holds only the Web pages that is more likely to be revisited. Whether visited Web pages are more likely to be revisited or not is determined by the characteristics of user's browsing behaviors. Our results clarify users' browsing characteristics make a filtering system of Web history work well. This study shows a solution of revisit Web pages, using Web history function.

Keywords: Web history · Revisitation · Refinding · Browsing characteristic · Filtering

1 Introduction

Now, using the Internet for information gathering has become very common. In information retrieval using the Internet, user often wants to revisit the Web page that was viewed in the past. In these cases, most people try to trace similar or same routes in their previous visit (e.g., use a search engine, follow a link). However, when people don't remember the page title, the search query, or the route of the web surfing, they can't sometimes reach their previous visited pages. In such case, there is a way to use a Web history function. However, there are some problems in current Web histories and few people use it to revisit Web pages. For example, most Web histories are displayed in text format. As the visited Web pages increase, a lot of efforts are required to find target Web pages from a lot of list.

There are some sorts of situations when it is difficult for the user to revisit target Web page using Web history function. In one situation, there are many similar Web pages in the Web history because of visiting many Web pages to achieve an information retrieval task. In another situation, user cannot recall the pages because he/she does not visit them for a long time.

In this study, especially, to decrease useless information on the Web history, we propose a filtering system that holds only the Web pages that is more likely to be

© Springer International Publishing Switzerland 2015
S. Yamamoto (Ed.): HIMI 2015, Part I, LNCS 9172, pp. 273–279, 2015.
DOI: 10.1007/978-3-319-20612-7_26

revisited. Since the Web page that user felt interesting is more likely to be revisited, we proposed a filtering system that judge whether the visited Web pages are beneficial for users' tasks from their behaviors on the visited pages (we call this "browsing characteristics"), and hold only the pages that will seemingly revisited.

2 Related Work

A number of studies have been studied about Web browser so far.

Matsuo et al. [1] studied about browsing history. They proposed a keyword extraction method in consideration of the user's interested fields using browsing history. In this study, they tried to identify user's interest and focuses from individual browsing history. Hijikata et al. [2] studied about users' characteristics in browsing. They proposed a keyword extraction method to estimate the user's interests using mouse operations. Minami et al. [3] proposed a Web page filtering system in searching fit for the user's interests. They estimate user's search intention using his/her browsing characteristics and the natural language processing to the visited pages. We applied some parts of their methods to the Web history function. Sungjoon Won et al. [4] studied to improve usability of the Web history function. They improved it using visual and contextual cues. However, they didn't improve a fundamental factor of information overload. In our study, we approached this matter by a different way of removing the useless web pages from the history.

3 System

3.1 System Outline

Our proposed system filter out only the Web pages the user is not interested in from the Web history and holds only the pages that he/she visited and were interested in based on the collected behavior on each page in user's browsing. Detail of filtering module is described below.

3.2 Filtering Module

In filtering module, the system passes through only Web pages which users were interested in. Whether the visited Web pages are interesting or not for the user is determined by the browsing characteristic he/she behaves on each Web page he/she visited. Table 1 shows the browsing characteristics used for filtering.

Table 1. Recorded browsing characteristics.

Browsing characteristics	Details
Slight mouse movement	It count mouse moving; more than 10 pixel per 0.1 s
Sorjoum time (s)	Time between page visit and page leave
Page text	Text that deleted tag data from HTML

The filtering system we proposed consists of two steps. In the first step, the sojourn time on each visited page is used to filter. The purpose of first step is filtering the Web page whose contents the user is not interested in at all. The threshold of the sojourn time used in the first step will be determined in the experiment below. Here, when user visits the same page multiple times, the total sojourn time of each visit are used.

In the second step, whether the visited Web pages hold or not is decided by the frequency of the slight mouse movements or sojourn time per text. Frequency of mouse movement and sojourn time per text are defined as follows.

$$\text{frequency of slight mouse movement} = \frac{\text{amount of slight mouse movement}}{\text{amount of Web page text}} \quad (1)$$

$$\text{sojourn time per text} = \frac{\text{amount of the sojourn time}}{\text{amout of Web page text}} \quad (2)$$

The purpose of second step filters the Web page whose contents the user once feels interesting but does not judge beneficial. The frequency of slight mouse movement must be high when the user feels the page contents beneficial. Furthermore, the sojourn time per text can be considered to increase when the user feels it is beneficial for him/herself he/she must read it carefully to understand it. The thresholds of the frequency of the slight mouse movement and the sojourn time per text also determined in the experiment below.

4 Experiment

4.1 Methods

We conducted the experiment to evaluate the filtering system.

The subject asked to perform three information retrieval tasks. Detail of each task is shown in Table 2. To simulate a realistic information gathering tasks, we made tasks by referring to the question that is really asked in the Japanese positing Q&A site (Yahoo! Chiebukuro [6]). We selected the tasks which the subjects should visit many pages to complete the task.

When the subjects completed each task, they were asked to make out a report about an approximately 1,500 Japanese characters, which is filled out about one page of A4 sheet with a word processor [7], during the experiment, the subjects prohibited taking notes not to memorize the Web page contents that visit in the past.

Table 2. Tasks.

Task number	Task details
Question 1	Gather information of japanese declining industry
Question 2	Gather information of life in medieval Europian
Question 3	Gather information of industry in Africa

After the completion of each task, the subjects were asked to evaluate all pages they visited whether they are useful or not to make out the report. This result is used to evaluate our system by comparing with the results of our filtering system as the following values:

Filter validity: percentage of correct answers of filtering.

$$\text{Filter validity} = \frac{\text{current classification of filtering}}{\text{amount of visited Web page}} \tag{3}$$

Page recall ratio: ratio of the Web page that user may be re-display passes through the filter among the Web page that user may be re-display.

$$\text{Page recall ratio} = \frac{\text{current Web pages that pass through the filter}}{\text{amount of visited Web page}} \tag{4}$$

Page precision ratio: ratio of the Web page that user may be re-display passes through the filter among the Web page that passes through the filter.

$$\text{page precision ratio} = \frac{\text{current Web pages that pass through the filter}}{\text{amount of beneficial web page}} \tag{5}$$

F-measure: comprehensive measure in consideration of recall and precision.

$$\text{F} - \text{measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{6}$$

Ten people were involved in this experiment, and they were from 20 to 23 years old. Though we didn't instruct to use a specific search engine to do information retrieval tasks, all subjects use Google search engine [5]. In addition, after the experiment, they were asked to fill out the paper questionnaires which contain the questions about their daily use manner of Web browsing such as usage of Web history, and their impressions of the experiment, such as the difficulty of each task. Looking at this questionnaire, all of the subjects accessed the Internet in everyday life and there is no problem in operating the Web browser. Furthermore, no people usually use Web history function to revisit Web page, which we reconfirmed that there are usability problems with current browser history.

The questionnaire also says that seven subjects said there is no difference between the questions difficulty. Only two subjects said question 1 is a little more difficult than the other questions and only one subject said question 3 is more difficult than the other questions. Thus, we consider that three questions are no difference about the difficulties among the tasks.

The Web browser used in this experiment includes only the basic functions of popular Web browsers. For example, back button, forward button, URL bar, and display area (Fig. 1). This browser was built on C#.

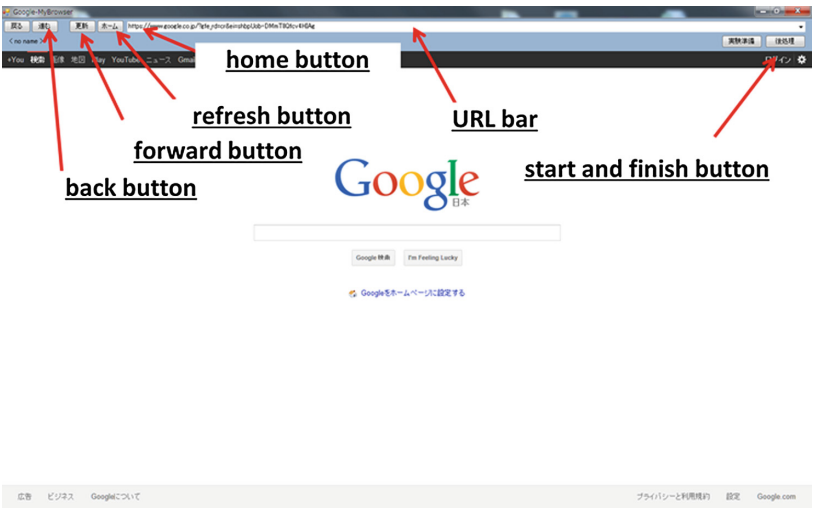

Fig. 1. Screenshot experimental Web browser.

5 The Results

5.1 Determination of the Thresholds of Each Filter

We determined the threshold of each filter by the result of this experiment.

In the experiment, amount of beneficial Web pages increased when their sojourn time was more than 10 s. Thus, we considered that threshold of sojourn time used in the first step should be set to 10 s. In the same way, frequency of slight mouse movement is set to 0.001 and sojourn time per text is set to 0.003 s.

5.2 Evaluation of the Filtering

In this experiment, the subjects visited 240 pages in total and felt 156 pages are interesting and 84 pages are not interesting among them. Our proposed filtering system filtered 47 pages, and 14 Web pages were mis-filtered among them.

Table 3 shows the result of each evaluation value.

In Table 3, the recall rate and F-measure were good. Therefore, it can be said that the filter works well overall.

However, several mis-filterings were occurred. In this experiment, 11 cases of 14 mis-filterings were occurred in the first step. Table 4 shows the number of mis-filterings in the first step for each subject

Looking at Table 4, most mis-filterings in the first step show a concentration in some subjects, such that 5 mis-filterings were occurred in one subject. In addition, the features of the sojourn time were also different by each subject ($M = 31.70$, $SD = 7.88$). This means that we must set up the thresholds considering the individual differences of browsing characteristic of each user.

Table 3. Result of evaluation value.

Validity	Recall	Precision	F-measure
0.729	0.910	0.740	0.814

Table 4. Relationship between the subject and mis-filtering of first step.

Subject	a	b	c	d	e	f	g	h	i	j
mis-filtering of first step	0	5	0	0	0	0	1	2	1	2

6 System Improvement

As explained above, we need to improve the first step filtering and to take individual differences of browsing characteristics in account. Therefore, to make the filter of the first step and to support the individual differences of browsing habits, we set the threshold dynamically changed according to each individual's browsing characteristics. We defined it as below:

$$\text{Individual variable threshold} = \frac{\text{individual average sojourn time}}{\text{n}} \tag{7}$$

Here, "n" was a fixed number, and we set it 3.0 here.

Table 5 shows the result of the improved system and Table 6 shows the number of mis-filtering of first step in each subject.

Showing Tables 3 and 5, all evaluate values was improved. In comparison between Tables 4 and 6, total mis-filtering and the difference among the subjects were reduced. Therefore, it is effective to automatically adjust the values for filtering to individuals according to their browsing habits.

Table 7 shows a result of first step mis-filtering with improved system.

Table 5. Result of filter restricting.

Validity	Recall	Precision	F-measure
0.754	0.917	0.757	0.830

Table 6. Relationship between subject and mis-filtering of first step of restricted filter

Subject	a	b	c	d	e	f	g	h	i	j
mis-filtering	0	3	0	0	0	0	1	3	1	3

Table 7. First step mis-filtering with improved system.

Task	First step mis-filtering
Question 1	2
Question 2	0
Question 3	8

Table 8. Result of visited Web pages for each question.

Task	Visited Web pages (±s.d.)
Question 1	7.9 ± 3.2
Question 2	6.8 ± 2.3
Question 3	9.3 ± 4.9

From Table 7, most of mis-filterings in first step were happened in question 3 and Table 8 shows the result of the number of each subject's visited Web pages for each question.

From Tables 7 and 8, the mis-filtering in first step tend to increase as the standard deviation of the numbers of the visited Web pages increased. As we do not clarify the cause of this problem, to solve it is one of our future works.

7 Conclusion and Future Work

Since current Web history functions include much useless information, few people use it to revisit Web pages. In this study, we suggested filtering system of Web history that passes through only beneficial Web pages.

In conclusion, we suggested a filtering system of Web history function using users' browsing characteristics. This result showed a method that can decide the benefit of visited Web page from the browsing characteristic he/she behaves at the Web page and it is effective to automatically adjust the values for filtering to individuals according to their browsing habits.

However, the mis-filterings in the first step were happened when the numbers of the visited pages vary widely among the users. In the future, we should clarify the cause and overcome this problem.

References

1. Matsuo, M., Hayato, F., Mitsuru, I.: Browsing Suppot by Highlighting Keywords based on User's Browsing History, vol. 101, pp. 85–92. Technical Report, The Institute of Electronics, Information and Communication Engineers (IEICE) (2002)
2. Yoshinori, H., Yoshinori, A., Younosuke, F., Amane, N.: Text Part Extraction Based on Mouse Operations and Evaluation of Extracted Keywords, vol. 43, pp. 566–576. Information Processing Society of Japan (IPSJ) (2002)
3. Shoutarou, M., Makoto, O.: Extraction of Search Intention based on User Behavior and Classification of Search Result, vol. 8, pp. 1–6. Information Processing Society of Japan (IPSJ), HCI (2011)
4. Sungjoon, W., Jing, J., Jason, I.S.: Contextual web history: using visual and contextual cues to improve web browser history. ACM Conference on Human Factors in Computing Systems (CHI) (2009)
5. Google search engine. https://www.google.co.jp/
6. Yahoo! Chiebukuro. http://chiebukuro.yahoo.co.jp/
7. A document format of Electronic official document. http://www.soumu.go.jp/main_sosiki/gyoukan/kanri/dtd01.htm#04