Personal Computer Operation History Data Collection System—Memory Retriever

Akimichi Tanaka† *and Tadasu Uchiyama*

Abstract

In this article, we introduce our system called Memory Retriever, which automatically collects the history of a user's operations on a personal computer and explain its functionalities.

1. Introduction

With advanced broadband access services becoming more widespread, people are spending more time on computers searching for information, shopping for personal goods, and engaging in other activities. The history of a user's computer operations is a valuable source of information because it reflects his or her likes and tastes. However, most data of this type is currently not used, but just wasted. Therefore, we have developed a system, called Memory Retriever, that automatically collects and accumulates personal computer (PC) operation histories for subsequent use in other advanced services. Some usage examples of users' PC operation histories and those of their colleagues and third parties are given in **Table 1**. Each history can be used for many purposes.

2. Requirements

We considered whether or not to require the installation of an application on each user's computer. Without a dedicated application, only a limited set of historical data such as Internet accesses can be acquired from web server logs or proxy server logs, as indicated in **Table 2**. To allow more detailed information to be obtained, we chose the approach that does require the installation of an application on the user's computer because we think that the advantages outweigh the disadvantages.

3. System architecture

An overview of our system is given in **Fig. 1**. The system basically consists of the client PC (C-PC) and the history collection server (HCS). The former is currently implemented using Windows XP or Windows Vista. The latter runs Cent OS.

C-PC captures the operation history and sends the data to HCS, which includes the following functionalities.

(1) Browser add-on:

It retrieves the browsing history via a web browser add-on. We have developed plug-ins for Internet Explorer and Firefox.

(2) History retrieval module:

This module consists of a history retrieval plug-in and a history collection program. The former captures the user's operation history and writes the data into operation log-files. The latter accesses and processes the data written in the operation log-files and stores the results in the operation history database. The plug-in has a structure that simplifies the handling of additional user input/output devices such as touch-pads. The browser plug-in receives data retrieved by the browser add-on and writes the received data in operation log-files.

† NTT Cyber Solutions Laboratories
Yokosuka-shi, 239-0847 Japan

Table 1. History usage examples.

History	Usage examples
User's own history	A user can analyze his or her own computer behavior by reviewing computer operations objectively.
Colleagues' histories	Employees can avoid redundant searches and improve their work efficiency by sharing online search activities.
Third parties' histories	Trends of various groups can be defined by categorizing histories by age and gender,

Table 2. History collection methods.

Method	Browsing web pages		Operations other than web browsing	Easy PC setting	Load on PC
	URL retrieval	HTML retrieval			
PC application	Good	Good	Good	Poor	Poor
Web server log	Poor (good only for own website)	Poor (good only for own website)	Poor	Good	Good
Proxy server log	Good	Good (not available on http websites)	Poor	Good	Good

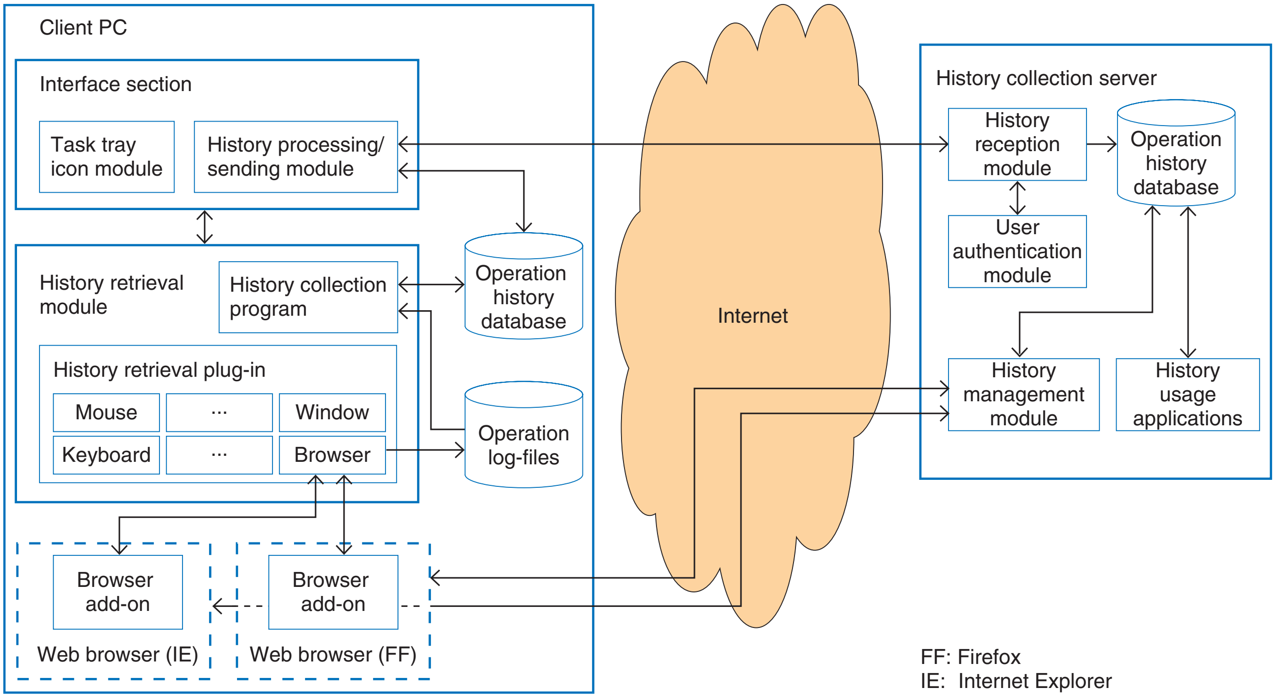


Fig. 1. System architecture.

(3) Interface section:

This section consists of the task-tray icon module and the history processing/sending module. The task-tray icon module shows icons that are always in the task-tray and acts as the starting point of the interface with the user. The history processing/sending module

Table 3. Retrievable history data.

Plug-in	Retrievable data
Mouse	Double click, button down/up
Keyboard	Key down
Window	Title, application name, name of file in use, thumbnail image
Browser	URL, HTML file, thumbnail image, title, referrer, search word, anchor text
Clipboard	Copy, paste
Printer	Start, end
File	Copy, move, erase

reads data from the operation history database and sends it to the history collection server. The module also extracts keywords used on search engine websites such as "goo" as well as the anchor text of clicked links. Data that has already been processed is erased after a preset time.

HCS receives the operation history sent via C-PC and accumulates the data in its database. It has the following structure.

(1) History reception module:

This section receives the operation history from C-PC and accumulates it in its operation history database.

(2) Operation history database:

This stores the operation history data sent via C-PC.

(3) User authentication module:

This module checks user names and passwords to ensure that a recently received operation history is associated with the correct user.

(4) History management module:

This module manages data in the operation history database.

(5) History usage applications:

Applications will be created to satisfy users' demands for how they want the acquired history to be used.

4. Retrievable data

Currently available plug-ins for operation history retrieval cover the mouse, keyboard, windows, browsers, clipboard, printers, and file access. History data that can typically be acquired by these plug-ins is listed in **Table 3**. Each data item is timestamped. Users can access the history management module on HCS via an ordinary web browser. An example of a web browsing history is shown in **Fig. 2**. Each web browsing sequence is shown as one unit. The top row is the history of the actions performed between 16:54:08 and 17:18:21 on April 9, 2010. Thumbnails of web pages visited during the time period are shown (partially). Users can access detailed information about each page by clicking one of the time periods shown.

5. Unique functions

Our system has the following unique functions.

(1) Scalable history retrieval

Our system can easily handle additional devices whose history is to be made available for retrieval if a history retrieval plug-in is created and placed it in a folder.

(2) Detailed information about web browsing

It can acquire not only URLs (uniform resource locators), but also HTML (hypertext markup language) files, thumbnails, and HTML files of small frames on frame pages. It can also retrieve search words as well as anchor text.

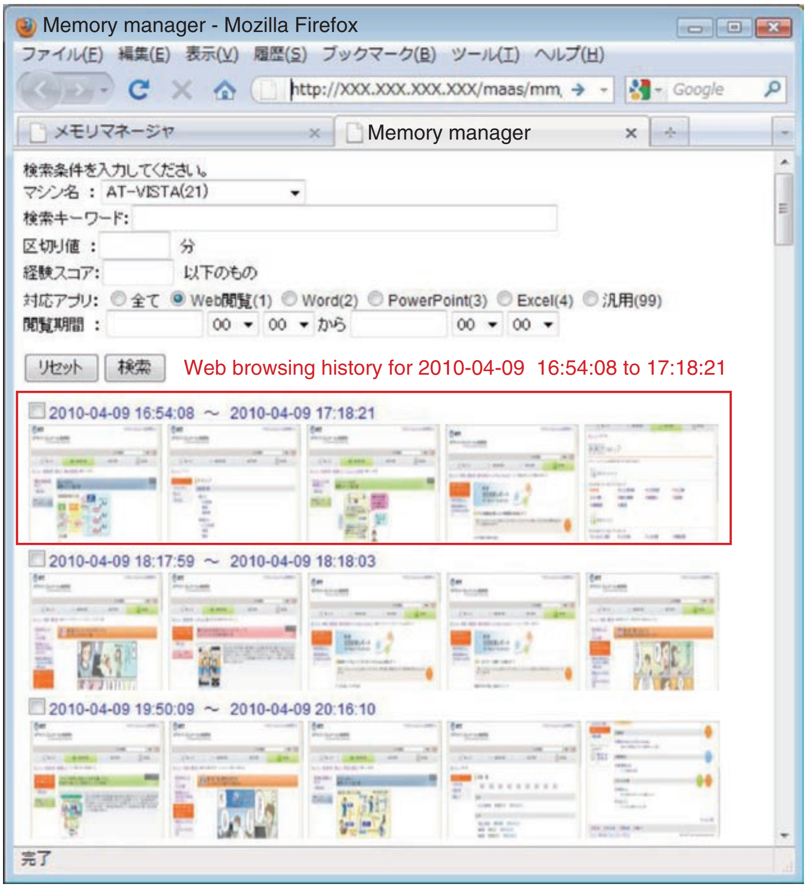
(3) Privacy protection

Since web browsing information may include users' private information, two functions are provided to protect their privacy.

- Protected-word anonymization function:

When words present in a protected-word list are included in a received HTML file, they are anonymized by being converted into XXX before being sent to HCS.

- URL filter function:


Memory manager - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://XXX.XXX.XXX.XXX/maas/mm
Google
メモリマネージャ
Memory manager
リセット
検索
完了

検索条件を入力してください。
マシン名 ： AT-VISTA(21)
検索キーワード:
区切り値 ： 分
経験スコア: 以下のもの
対応アプリ: ○ 全て ● Web閲覧(1) ○ Word(2) ○ PowerPoint(3) ○ Excel(4) ○ 汎用(99)
閲覧期間 ： 00 00 から 00 00

Web browsing history for 2010-04-09 16:54:08 to 17:18:21

□ 2010-04-09 16:54:08 ～ 2010-04-09 17:18:21
□ 2010-04-09 18:17:59 ～ 2010-04-09 18:18:03
□ 2010-04-09 19:50:09 ～ 2010-04-09 20:16:10

Fig. 2. Example of web browsing history.

A white list and a black list are provided. History data is stored when users are browsing URLs on the white list but not when they are browsing ones on the black list.

(4) Greater amount of retrievable window information

As Table 3 shows, the name of each application and file used can be retrieved via a window plug-in. Although it does not show detailed information about each application, it shows rough information, such as file opening/closing times, which means that most PC operations can be understood.

6. Advanced usages

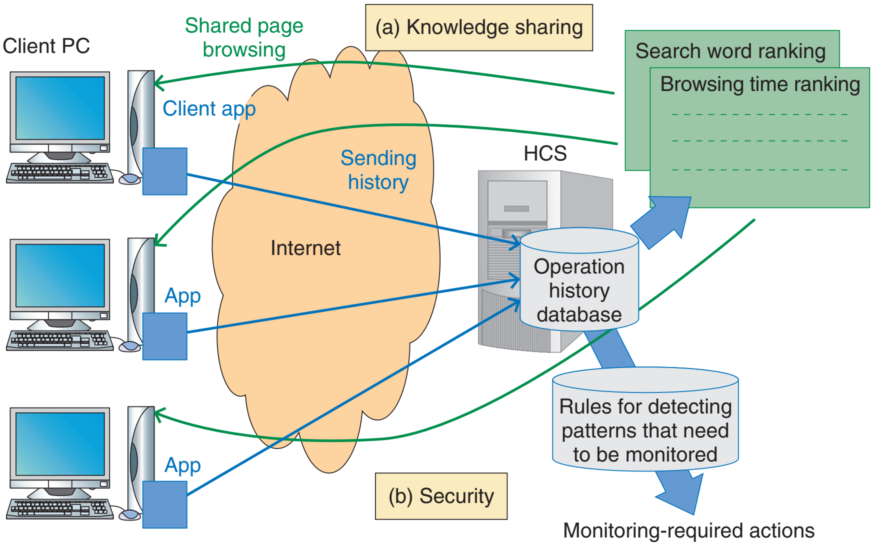
Below we give three examples of how retrieved history data can be used.

(1) Own behavior review

A user can understand his or her computer behavior objectively, as shown in Fig. 2, by reviewing the PC's web browsing operation history. As a result, the user may realize that he or she is spending too much time on a particular web page or has stopped visiting certain websites, which might lead to behavior changes for better computer use.

(2) Knowledge sharing at the office

Since operating histories are collected on HCS, office workers might feel uncomfortable about all the web pages that they have visited being logged in one location. Instead of identifying web pages individually, it might be better to disclose a single web page that visualizes only statistical information about the computer operations of all the workers. Such a web page could include ranked lists of the most-used search words or web pages where people stayed for the longest time. The usage flow for data sharing is



App: application

Fig. 3. Advanced usage.

shown in (a) in **Fig. 3**. The sharing of such knowledge at the office will help raise the efficiency of work.

(3) Security

In recent years, web-based computer systems have becoming common at many companies and municipalities across Japan, and browser histories are becoming more valuable as data sources. Our system allows a warning module to be implemented that will help prevent information leakage. This module will use predefined monitoring patterns to detect inappropriate operations such as company files being accessed and forwarded by web mail. The usage flow for security is shown in (b) in Fig. 3.

7. Future plans

We will continue researching ways of increasing the number of items whose history can be retrieved, including detailed operation information about popular Microsoft Office applications (Word, Excel, and PowerPoint), email, and USB (universal serial bus) memory. Considering that it may become difficult to display one complete history record when there are many items, we are developing a scoring method that will display only high-scoring items. We also plan to develop history usage applications that satisfy users' needs.

Akimichi Tanaka

Senior Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.E. and M.E. degrees in precision machinery engineering from the University of Tokyo in 1985 and 1987, respectively. He joined NTT Information Processing Laboratories in 1987 and has been engaged in research on artificial intelligence, pattern recognition, and learning support systems. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan and the Information Processing Society of Japan (IPSJ).

Tadasu Uchiyama

Senior Research Engineer, Supervisor, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.S. and M.S. degrees in physics from Nagoya University, Aichi, in 1985 and 1987, respectively. He is interested in web services and technologies. He is a member of IPSJ and the Japan Society for Industrial and Applied Mathematics.