31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

The Exploration & Forensic Analysis of Computer Usage Data in the Elderly

William J. Hatt, Edward A. VanBaak, Holly B. Jimison, *Member IEEE*, Stuart Hagler, Tamara L. Hayes, *Member IEEE*, Misha Pavel, *Member IEEE,* Jeffery Kaye, *Member IEEE*

Abstract—**Unobtrusive in-home computer monitoring could one day be used to deliver cost-effective diagnostic information about the cognitive abilities of the elderly. This could allow for early detection of cognitive impairment and would additionally be coupled with the cost advantages that are associated with a semi-automated system. Before using the computer usage data to draw conclusions about the participants, we first needed to investigate the nature of the data that was collected. This paper represents a forensics style analysis of the computer usage data that is being collected as part of a larger study of cognitive decline, and focuses on the isolation and removal of non user-generated activities that were recorded by our computer monitoring software (CMS).**

I. Introduction

With the growing elderly population in the United States and around the world[1] there will be an increased need for preventative care for the aging population. This need stems both from our duty to provide them with increased quality of life[2], and the financial realities that we will face when caring for a rapidly growing population.[3, 4] Previous studies[5-7] have shown the importance of early recognition of cognitive decline, but current tests are expensive, time consuming, and are administered infrequently.

The Biomedical Research Partnership (BRP) is a NIA funded longitudinal research study involving over 230 elderly participants. The study focuses on continuous and unobtrusive in-home assessment of physical activity and computer usage.[8] One of the aims of the study is to determine whether the unobtrusive monitoring of general activity in the home can be used to detect changes in motor and cognitive function, thereby allowing for early intervention and an increased quality of life. Physical activity is being monitored through motion sensors, and computer usage is being monitored through a program that is installed on participant's personal computers.

The benefit of the approach used in the BRP study, compared to traditional methods, is that the monitoring is continuous. The cost of the testing also drops as specially trained individuals are not required at the home in order to administer tests or monitor the situation.

Of the more than 230 participants in the study 189 of them have Internet connected computers that are monitored as a part of the study. While the scale of the project promises to offer large quantities of useful information that can be used for later research it also poses a challenge to our ability to carefully manage and process the computer usage information. Additionally, we needed to control for several types of computer events that could lead to a misinterpretation of valid computer activity. This was due to the necessary but often unpredictable nature of the interactions taking place between the systems involved, the users, the operating system, and the other applications running on the computer. We found that a computer forensics perspective should be taken with the computer usage data in order to ensure its quality and accuracy. The term *computer forensics* is meant to convey that before looking at the data in order to draw conclusions about the participants we first needed to investigate the nature of the computer data that was collected. This paper represents a analysis of the computer collection techniques that are being used in the BRP computer usage study.

II. Methods

Assessment Setup:

The Computer Monitoring Software (CMS) that was used in the study was designed at OHSU's Division of Biomedical Engineering (BME). The software functions in two ways. Before the user logs onto the computer, the CMS prompts the users to enter a user name and password given to them at the start of the study. This prompt screen is actually a locked screensaver that acts as a replacement for the standard windows login screen. During and after the entry of a user name / password combination, the CMS monitors the activity of the user.

Manuscript received April 7, 2009. This work was supported by the following National Institute on Aging (NIA) Grants; Oregon Roybal Center for Aging and Technology P30AG024978 – (for use of ORCATECH pilot grants, living laboratory, and/or administrative core support); Oregon Alzheimer's Disease Center (OADC) P30AG08017 – (for use of BRP/ORCATECH clinical data); Bioengineering Research Partnership, Oregon Roybal Center for Aging and Technology R01AG024059 – (for use of BRP data)

W. J. Hatt, E. A. VanBaak, and H. B. Jimison are with the Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239-3098 USA (e-mail: hattb@ohsu.edu; vanbaake@ohsu.edu; jimisonh@ohsu.edu).

M. Pavel, T. L. Hayes, S. Hagler are with the Division of Biomedical Engineering, Oregon Health and Science University, Portland, OR 97239-3098 USA (e-mail: pavel@bme.ogi.edu; hayest@bme.ogi.edu; haglers@bme.ogi.edu).

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

Five types of events were recorded from the participants' computers (Trigrams of general typing data, Login events, Login Passwords – a.k.a. KeyData, Application focus change events, and Mouse events). Because of storage constraints each type of data has to be stored separately, and then later pooled in a single location so that each type of event could be viewed in relation to the others and ordered by date. Several graphing techniques were adapted for use in order to explore the data and look for errors.

The host computers were uniformly equipped with Microsoft Windows XP, though hardware and peripherals such as mouse input devices were not as uniform. A primary concern of this investigation was to discover and resolve and unexpected behavior that may occur between the operating system and the peripheral hardware. This was done to ensure that the information being gathered from the study participants could later be correctly interpreted.

The login event is stored, as well as all of the keystrokes that were used to generate the password, and recorded with millisecond precision. Once the user has logged in, the program also monitored various types of mouse input, such as movement and mouse click events. We also monitored application activity and recorded the path and time at which an application window gained focus on the user's screen.

Although keyboard typing activity is recorded, this information is restricted and obfuscated over concerns for the patient's privacy. Rather than recording actual keystrokes in the order they were entered, the keyboard data is recorded in the form of trigrams. Trigrams consist of the last 3 key codes that the user has entered, and the amount of time that has elapsed between the 3 key press events. The trigrams are also time stamped with a limited 1 hour resolution, which allows later researcher to have information about the approximate time the key information and how fast it was entered without the ability to recreate what was actually typed. Furthermore, key activity is only recorded when the participant is using a web browser or typing an e-mail in Microsoft Outlook.

The information that was collected by the CMS was forwarded to a special purpose computer stored at the participant's house. Once a day all of this information is bundled and forwarded over a broadband Internet connection to a secure MySQL database server that is run at OHSU's Division of Biomedical Engineering (BME). Due to the quantity of data that is generated and the fact that much of the data will be used for longitudinal investigations, the design of the back end system that we used is subject to a complex set of requirements. Each data type is stored in its own table, with the exception of the mouse data which because of its size is stored in a separate SQL server at BME. Although this allowed us to work with very large data sets, and it scales well as the number of study participants increases, it also presents a computational challenge. Our current investigations, and undoubtedly later analyses, will require that all of the information for a specific user be collated into a format for easy and efficient access.

We used TheMathWorks^{TM} MatLab 2008b software to process the user's computer activity data. This was accomplished by connecting MatLab to the MySQL database and pooling the data into a single set of files for each user. The sheer volume of data that is collected for each user made it necessary to store user data into separate files, where each file contained data from a single month. Multiple months can be loaded into memory as long as enough RAM is available. This approach satisfies the memory limitations that we faced, while allowing investigators to choose the quantity and duration of the information needed. Changes in MatLab 2008b include a greater focus on object oriented programming which allowed for rapid algorithm development. The graphing functions in particular were a valuable tool that was used in our investigation of the data that was being collected.

Confounding Factors:

During initial analysis it became apparent that memory management was a major road block standing in our way of processing the data. While our aim is to compare data from distant times in the study, we were limited by our ability to load all of the information into memory or in some cases even on to a single computer. The SQL databases were queried from MatLab and prepared for processing by storing the collated user data in a separate file for each month. Once prepared in this fashion the data was loaded into memory in discrete chunks and we were able process one or several months of data at a time. This allowed for old data to be archived as ready for processing, and for the new data from continuous monitoring to be updated as needed.

Continuous monitoring coupled with a longitudinal study meant that there were large volumes of data that needed to be stored and processed. The mouse data in particular has generated over 8GiB for two years of data for a single user. Multiplying that storage demand for all 196 users, and the number of years left in the study and it becomes apparent that the growing size of the SQL database presents a challenge to both our ability to store and process all of the data.

Collation of the data sources that were recorded (Trigrams, Login Events, Passwords, Applications, Mouse Events) within MatLab allowed for each type of information to be viewed in relation to the other types and ordered by date. Because of memory limitations, only the type of event and the time it occurred are stored in memory. Additional information about the event, such as mouse cursor position for mouse events, or application paths for application change events are loaded on a need-to-know basis from the SQL server in order to save memory. Several graphing techniques were adapted for use with our computer usage data set.

III. Results

We used polar plots of computer events to visually explore user activity throughout the day. Examples of polar plots are shown in Figures 1 and 3. The polar plots show activity of the user by time of day represented by the angle of the polar plot, with each successive ring around the circle representing another day. Viewing the data in this fashion became a valuable diagnostic tool that was used to help ensure the validity of the data that was being collected.

Mouse Data Noise Reduction:

One trend the polar plot was able to reveal was that not all of the mouse data was generated by the user. We found that the CMS generated mouse movement data in a "heartbeat" fashion, approximately every 30 minutes. This happened because the monitoring software was designed to empty its memory at regular intervals in order to prevent buffer overflow and data corruption. The problem was that not all of the mouse data was actually removed from memory and this resulted in a single time-stamp of mouse activity occurring approximately every 30 minutes (See Figure 1) and it needed to be filtered out before other teams began analyzing the mouse data.

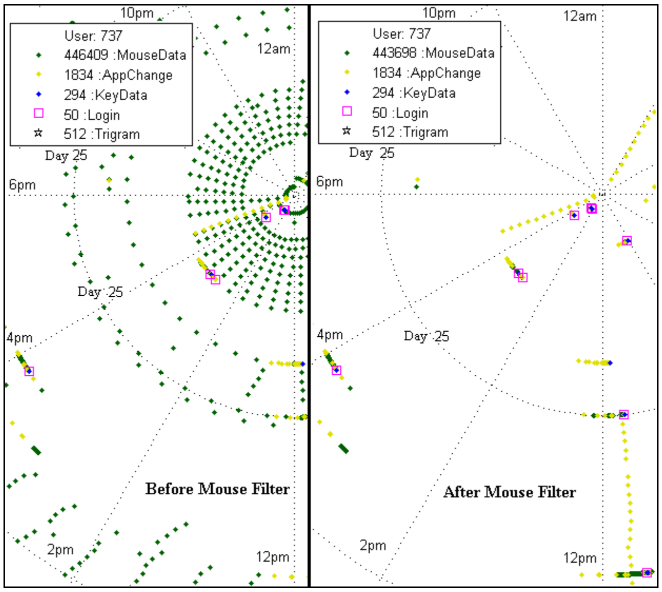

Figure 1: Categorical data plot and mouse filter: Spokes radiating from the center indicate the time of day. Each successive ring away from center represents a different day, and the number of data points for each type is listed to the left of the name in the legend. Left Side: Before the mouse noise filter; Right Side: After the mouse noise filter. *View of graph is zoomed in; not all data points are shown.*

The short streaks/lines in Figure 1 represent actual computer use and are a mix several categories of data. They are not actually lines, but tightly packed with many individual data points. In contrast the lone mouse data points represent the heartbeat pattern that was generated by the CMS. Filtering of the mouse data was possible because of the density of the mouse data points. During normal use, the mouse generated a new time stamp every 15-300ms depending on how far or fast the mouse was moved. In contrast, the noise generated, by our CMS occurred approximately once every half hour and generated only a single data point. In practice the user could not generate a single time stamp by moving the mouse, because even a small mouse movement by the user would trigger the computer to record several data points.

Due to the high density of data that was collected from the participants' computers (particularly the mouse data) we were able to develop an alternative and more sensitive approach for determining the amount of time the users were spending on the computer. Rather than depending solely on the login and logout events (which may artificially increase apparent computer usage) we estimated computer usage by looking for large gaps in the timestamps between data points and removing them from our usage calculation. The remaining gaps represent the actual amount of time that was spent on the computer.

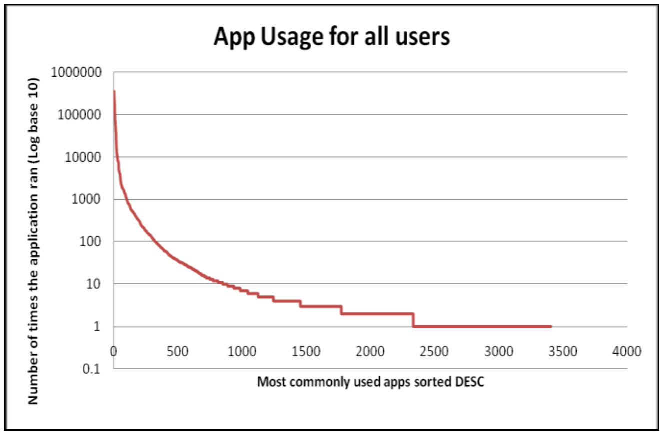

Figure 2: Application usage for all CMS users: The Y axis indicates the number of times an application was executed. X axis indicates the relative rank-number of each application *E.g. The #1 most used application was executed 386,642 times, while the least used (application numbers 2331 through 3401) were only executed once.*

Application Data Analysis:

Counting the number of unique applications that were used by all participants (see Figure 2) shows that some applications are used quite often, but most applications are used quite infrequently. Of the 3401 unique application paths that have been recorded, 24 of them account for more than 90% of all the applications change events. The top 282 programs account for 99% of the application usage, while the other 3119 applications account for less than 1% of the total. The exponential decay of the application usage made it a worthwhile approach to bin and sort the applications in order to better understand the computing habits of the participants.

When application change events were recorded in the database (see Figure 3) there was undoubtedly activity on the computer, but the caveat is that possibly not all of the activity was initiated by the user. Automatically updating programs such as antivirus software, Windows Update, and Windows Defender can present false positives for computer activity and decrease the accuracy of any estimates that are made concerning the amount of time participants are using their computer. Separating this kind of program activity from normal computer use can be difficult, because even though the programs are somewhat predictable because they auto update at regular intervals they can also be manually started by the user. Removing all instances of an auto executing program from our calculations is thus not an option, because it would also remove valid user activity. This kind of noise can, however, be corrected for by searching for all unique program paths are executed at nearly the exact same time each day and then to remove them from the data set. This is an advantageous approach as it does not depend on identifying new software or maintaining a list of software that runs on a regular basis.

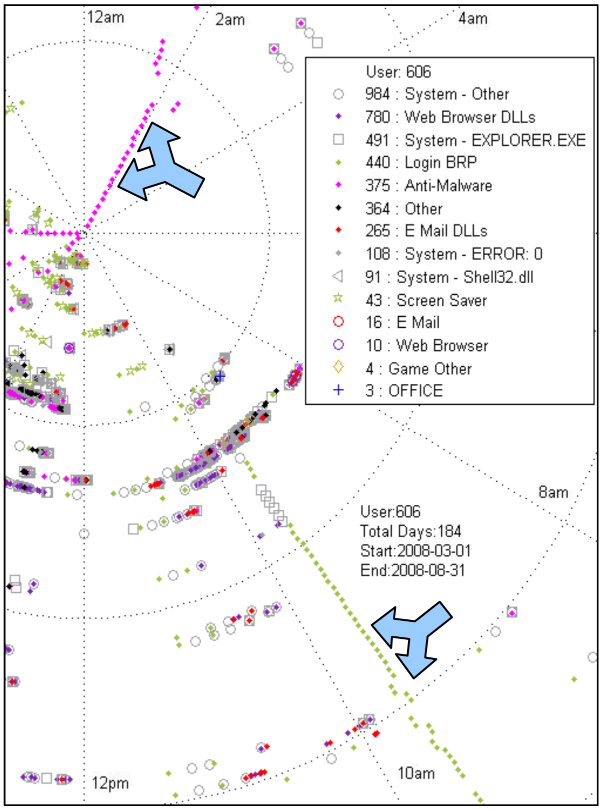

Figure 3: Application data polar plot: Upper left arrow (Anti-Malware) and lower right arrow (LoginBRP) indicate automated system functions that are not user generated. ****Not all 184 days are shown on the graph.*

IV. Discussion and Conclusion

The primary purpose of the data collection in the longitudinal BRP Project is the assessment of potential behavioral markers of cognitive and neuropsychological state of the participants. This objective requires a high degree of data integrity and interpretability. With the eventual goal of using the computer data in conjunction with non computer measures of activity and performance, we needed to remove obvious artifacts that were left from automatic system processes. We were able to clean the dataset so that it better represented human usage only, and thereby demonstrated the feasibility of monitoring the computer usage of elders in their home environment.

Had we not carried out this initial step of analysis, our computer usage calculations would have been faulty through our assumption that all of the activity being recorded on the computer was actually user generated. The noise that was present in the mouse data was later confirmed to have come from the CMS package through a review of the source code, and team members were made aware of the situation. Through this process we have discovered the importance of a careful interpretation of the data before it is summarized.

Acknowledgment

Foremost we would like to thank all of the elderly participants in our study. Without their cooperation and understanding none of this would be possible. Additionally we would like to thank all of the technical teams at OHSU's Division of Biomedical Engineering (BME), and the Oregon Center for Aging and Technology (ORCATECH) for designing and assembling the systems that were used in generating the dataset that this paper analyzes.

References

[1] Centers for Disease Control and Prevention. Trends in Aging — United States and Worldwide. MMWR 2003;52:102-106.

[2] Maciosek MV, Edwards NM, Coffield AB, Flottemesch TJ, Nelson WW, Goodman MJ, Rickey DA, Butani AB, Solberg LI. Priorities among effective clinical preventive services: methods. Am J Prev Med 2006; 31(1):90-96.

[3] National Commission on Prevention Priorities. Preventive Care: A National Profile on Use, Disparities, and Health Benefits. Partnership for Prevention, August 2007.

[4] National Commission on Prevention Priorities. Data Needed to Assess Use of High-Value Preventive Care: A Brief Report from the National Commission on Prevention Priorities. Partnership for Prevention, August 2007.

[5] P. Erickson, R. Wilson, and I. Shannon, Years of Healthy Life. Hyattsville,MD: National Center for Health Statistics, 1995, Statistical Notes, no. 7.

[6] L. Boise, D. L. Morgan, J. Kaye, and R. Camicioli, "Delays in the diagnosis of dementia: Perspectives of family caregivers," Amer. J. Alzheimer's Disease, vol. 14, no. 1, pp. 20–26, 1999.

[7] P. Glascock and D. M. Kutzik, "Behavioral telemedicine: A new approach to the continuous nonintrusive monitoring of activities of daily living," Telemedicine, vol. 6, no. 1, pp. 33–44, 2000.

[8] Kaye, J., Hayes, T., Zitzelberger, T., Yeargers, J., Pavel, M., Jimison, H., Larimer, N., Payne-Murphy, J., Earl, E., Wild, K., Boise, L., Williams, D., Lundell, J., and Dishman, E., Deploying wide-scale in-home assessment technology, in Technology and Aging: Selected papers from the 2007 International Conference on Technology and Aging, Mihailidis, A., Boger, J., Kautz, H., and Normie, L., Editors. 2008, IOS Press: Amsterdam, The Netherlands. p. 19-26.