Towards Accurate Measurement Of Computer Usage In a Rural Kiosk

Rajesh Veeraraghavan[1] Gauravdeep Singh[2] Bharathi Pitti[2]
Greg Smith[1] Brian Meyers[1] Kentaro Toyama[1]

[1] Microsoft Research
[2] Birla Institute of Technology and Science, Pilani

Abstract. Rural PC kiosks are increasingly seen as a tool for socio-economic development in developing countries. In order to make kiosks successful, it helps to understand patterns of usage in existing kiosks. Often, questionnaires or interviews are conducted to determine usage patterns, but self-reporting by subjects is notoriously inaccurate. In this paper, we present a tool that allows accurate measurement of when and how PCs in a kiosk are being used. We discuss how an existing tool has been adapted for easy data collection in rural kiosks and present evidence that even regular users of computers are poor at estimating their own usage statistics.

1 Introduction

Rural PC kiosks (henceforth referred to as kiosks) are being set up in developing countries as a tool for supporting socio-economic development [2][11]. Kiosks are essentially shared-access computers optionally connected to the Internet. These are used by poor rural villagers who usually cannot afford to own a personal computer in their homes. Typical applications that run in these kiosks are e-government services, agricultural price information, computer skills training, and telemedicine [11]. The applications and services that are offered in these kiosks vary depending on the agency that is running them as well as the location and demographics of the host village.

Despite interest in these projects there have been surprisingly few critical evaluations of their utility and impact [6]. Some studies show anecdotal evidence that kiosks do not achieve their original goals of social or economic development [4]. Hypotheses for these difficulties include lack of reliable connectivity[9], lack of power, lack of relevant applications[9], economies too small to sustain a connected PC[11] and government bureaucracy [9].

More rigorous evaluation of kiosk projects requires more rigorous evaluation methods. Some researchers, in a bid to gather good data have begun surveys of kiosk operators and customers. Although surveys are an established way of gathering both quantitative and qualitative data, they have their drawbacks, chief among them that they rely on a subject's self-reported answers. In this paper, we describe a software-based logging tool we have developed that gathers precise usage statistics for rural kiosk PCs. We also establish the value of software-based logging tools by showing in a pilot test that people – even experienced computer users – are highly unreliable in estimating time spent engaging in different activities while at the computer.

2 Current Approaches to Evaluating Kiosk Projects

In the life cycle of a kiosk project, there are many instances when evaluation can help the project along [10]. In the initial stages, a needs assessment in the community, together with background demographic data may be helpful to determine what applications would be best run at the kiosk. During operation, usage statistics of the kiosk will be helpful in fine-tuning the kiosk's offerings. NGOs and others concerned with the social impact of a study will be interested in determining the impact of a kiosk on the community. And, those running kiosks as businesses will be interested in, for example, cash flow of the kiosk.

For each kind of evaluation, different methodologies are more suitable than others. To understand the local economics, for example, a visit to the district administrative office may be required. In gauging the interest of the community, an informal "focus group" with community members is useful. Interviews with kiosk users and operators can help glean useful information during the operation of the kiosk. More rigorous methods involve surveys and questionnaires administered to statistically meaningful populations of kiosk operators and customers.

For the purposes of understanding patterns of usage, most studies to date involve the use of survey instruments and questionnaires, where essentially a researcher visits kiosks and asks a predetermined set of multiple-choice or open-ended questions to kiosk operators, customers, and in some cases, non-customers [1]. These studies may also involve collection of data by examining books that the kiosk operator keeps to gather information, which provide information about, for example, how many users visit the kiosk.

Survey-based approaches are known to be only partially reliable, as they depend upon accurate self-reporting. In addition to factors of human error, operators, customers, and non-customers of kiosks may have incentives to misreport. For example, a kiosk operator may choose to over- or underreport income, in the hopes of gaining some favor with the kiosk agency (whose introductions are frequently necessary for a researcher to conduct the survey in the first place).

In addition, surveys containing technical concepts and jargon (e.g., ”How much do you browse the Internet?”) are often difficult to pose for the less computer-literate kiosk customers, who may very well be browsing, but not understand terms such as ”browser” or ”Internet.” Abstractions at this level may not be easily comprehended by the survey respondent [10].

There have been some attempts to actually study the users in the rural kiosks by physically observing how they use the computer [3]. Others have installed video cameras to watch children interact with a public computer kiosk [7]. These techniques have the drawback that they are immensely time- and resource-intensive, and require either active participation of the researcher or later coding of video. They cannot be easily scaled out or run over long periods.

Lately, researchers have begun to deploy software-based monitoring systems that, for example, log websites visited in a browser [7]. At one prominent rural-kiosk agency, all transactions happening over their web portal are logged on a back-end database [8].

3 VibeLog: Computer Logging Tool

We have adapted a software-based monitoring tool called *VibeLog* [5] to record computer activity at rural kiosks running Microsoft Windows. VibeLog was originally developed to better understand how multiple windows were managed in the Windows operating system (OS).

The Windows OS has a window manager which tracks all window activity in the user interface and raises events to public consumers when window changes occur. VibeLog takes advantage of this feature in Windows by programmatically hooking these public window event streams and recording them to file. This allows it to log all window-based activity on the machine without modifying the OS itself [5]. The main feature of VibeLog is the maintenance of two logs of window system information: events and windows. The log of events contains an entry for every window management activity reported by the OS (opening/closing/activating etc.), and the log of windows contains a series of entries enumerating the on-screen windows each minute that a user is active. Each log entry has a timestamp and contains window attributes and input information. Table 1 shows the raw data that the tool collects.

Serial No	Time Stamp	Window Activation	Window Name	Window Class	Process Name	Window Position
1	7312	NAME CHANGE	WFPLogs	ExploreWClass	Explorer.exe	0,0,1054,831
2	26815	DESTROY	Excel Worksheet	MS-SDIa	Excel.exe	0,994,1280,1024
3	107016	CREATION	Internet Explorer	IEFrame	IExplore.exe	22,29,982,751

Table 1. Sample VibeLog Event

The raw data is then uploaded to a database (SQL Server). This allows SQL queries to be run on the data based on the high-level questions we are interested in answering.

VibeLog was designed to collect data among American consumers who were running Windows at home with a reliable Internet connection. The tool was installed online over the Internet and was capable of updating itself by checking for newer versions online. VibeLog writes data locally to a PC as a simple text file, and sends the text file to a central server on the Internet upon start-up if the PC is connected.

3.1 Adaptations to the VibeLog Tool for the Kiosk Study

We made four kinds of changes to the original tool in order to adapt to for kiosk-usage monitoring.

First, the tool was augmented to collect additional data that the original VibeLog did not. Specifically, we added an option for the tool to log (1) a list of all Internet sites that were visited by a browser and (2) the hardware and software configuration of the machine.

Second, it was made more easily configurable to account for different privacy policies of kiosk agencies. We added fine-grain control by having the tool read a configuration file. The file is a simple text file which has boolean values for the configuration parameters. This can be changed at any point - but for the configuration changes to be taken into effect a restart is required.

And third, the installation and data-collection process were engineered to happen by the insertion of USB flash drives, since connectivity in rural kiosks is often poor or non-existent. The kiosk operator only has to insert a USB flash drive into a machine for installation to occur, and a separate USB flash drive is inserted for data collection (data is copied to the flash drive). We ensure that files from the same machine can be identified by appending a randomly generated, globally unique identifier to the data filenames. This also ensures that the same data-collecting USB drive can be used serially on many machines without file naming conflicts.

Finally, similar to the SQL server queries used by the original VibeLog tool, we have generated Excel macros that analyze the data for our purposes. For example, three of the analysis macros are as follows:

Email Time We considered both client based (Microsoft Outlook) and web-based email engines (Gmail, Yahoo mail, Hotmail). For Outlook, it looks for process name "OUTLOOK.EXE" and adds time spent in it. Also, if the process is "WINWORD.EXE" and the title is "*Message" or "*Message(*)" or "*Discussion", the time spent in this event is counted as mail time. We looked at the raw data to ensure that we weren't missing any other email clients.

Total Computer Usage We have time stamp data for every event (window-change related) as to when it happened. If between any 2 successive events the time gap is more than thirty minutes we mark the whole time as idle time. We keep adding this to the total idle time. Then in the event log file we subtract from the time-stamp of previous event the time-stamp of current event. From this we subtract the idle time to get computer usage. This is accumulated to get total computer usage for each user individually.

Number of Distinct URLs Visited We only look at the hostname part of the URLs to calculate uniqueness. For example, www.hotmail.com/content.html and www.hotmail.com/default.html will be only counted as one unique URL.

3.2 Privacy and Security Concerns

Any collection of user data must be done with privacy concerns in mind. Apart from legal waivers, signs posted to indicate that a logging tool is being used, and expectations that data collected will only be published in aggregate form, we also have made the following adaptations to the tool to ensure privacy:

- The configuration file gives flexibility to the kiosk operator to determine what gets logged. The kiosk operator can turn on/off logging of various information easily by editing a configuration file.
- The only identifying information that can be traced in any way to a specific computer or user are globally unique identifiers. While these ensure consistency of data between data-collection instances, they cannot be linked to a particular individual or PC.
- The content of the log files will be encrypted while stored on the machine.
- Data collection through the USB flash-drive insertion moves data files from the local machine to the USB, so that older logs of data are not stored indefinitely on the machine.

4 Experimental Study

4.1 Methodology Adopted for the Study

In order to test ease of use of the tool and to establish the value of the tool, we compared results of quantitative data collected with our tool with an electronic survey. We intentionally compared questions that could be answered by the tool which were questions asked in previous surveys on kiosks [11].

In this pilot study, volunteers were selected from colleagues at our research institution, all of whom were regular users of computers.

We had eight volunteers install VibeLog in their machines, using the installation USB drives. The participants were then left alone to do their normal work for two days. After two days we collected the data logs from their machines using the data-collection USB drive.

We then asked the participants to complete a survey about their activities while at the computer over the two days. The survey asked the following six questions:

1. How many Internet (web) searches did you conduct?
2. How many hours did you spend in e-mail applications?
3. What is the number of distinct applications you ran?
4. How many hours did you spend surfing the web?
5. How many hours did you spend on the computer cumulatively?
6. How many distinct Internet sites did you visit?

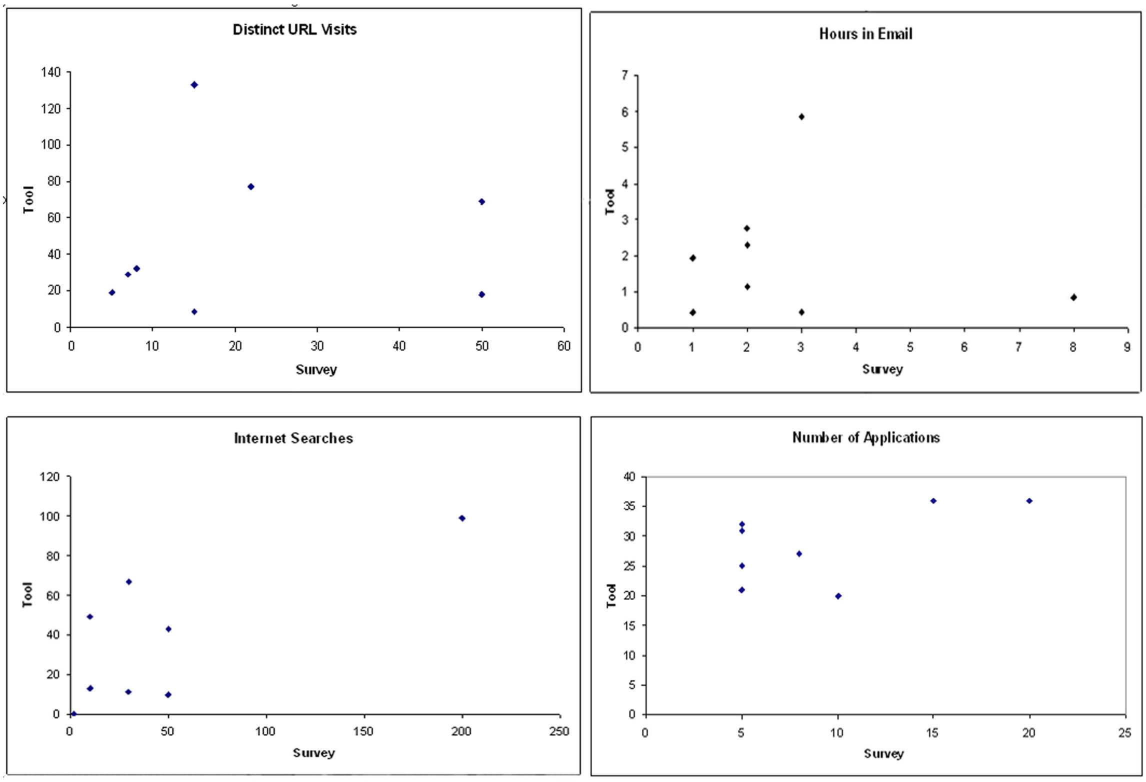

Fig. 1. Scatter plots of user reports versus data captured by the tool. The lack of correlation in the data shows how unreliable self-reports of PC usage habits are.

Question	Survey	VibeLog	Mean	Std Dev Diff
	Mean	Mean	Diff	Diff
Internet Searches	47.8	36.5	-11.2	-44.9
Email Time(hours)	2.8	2.0	-0.7	-3.0
Applications	9.1	28.5	19.3	-5.5
Internet Browsing Time(hours)	5.5	3.0	-2.4	-4.0
Total Computer usage(hours)	23.0	20.7	-2.3	-11.2
Distinct Urls	21.5	48.3	26.7	-44.5

Table 2. A comparison of data collected by the self-reporting and by the tool. The right two columns show the mean and standard deviation, respectively, of the differences between self-reporting and tool-captured data.

4.2 Analysis of Results

Table 2 shows the mean and standard deviation of the difference between what they reported and what was seen by the tool.

The VibeLog results were analyzed using Excel macros and answers to the questions asked in the survey were computed using data collected by the tool. The VibeLog results were plotted against the ones from the survey (Figure 1). Scatter plots are shown for four of the questions. If the reporting were 100% accurate, we would see these points lie on a straight line with a slope of one.

The reality is that the points are scattered with very low correlation coefficient. There is no easily explained pattern in how people mis-estimate their answers to the questions posed, although in some cases, there are consistent tendencies to under- or over-estimate.

We see, for example, respondents consistently underestimated the number of applications that they were using. On time spent in e-mail, half the respondents overestimated, while for the question of time spent browsing, respondents as many as 75% of the respondents overestimated.

In examining the variance of responses, we also see dramatic fluctuations, with many respondents being off by a factor of two in either direction for some questions.

5 Conclusions

We have presented a software-based logging tool specifically designed to gather usage statistics about PCs in rural kiosks. The tool captures information about URLs visited, time spent in certain classes of applications, amount of usage during the day, and so forth. Installation and data collection happens through USB flash drives, so that connectivity is not required (an earlier version of the tool is capable of recovering data over the Internet).

We also ran a pilot experiment with regular users of PCs to show that people consistently mis-estimate what activities and how much time they spend on a computer, indicating that a tool of this nature is necessary for accurate data collection of usage statistics.

Now that the tool has been designed and tested, in future work, we intend to deploy it for better understanding of rural kiosk PC usage.

References

1. An evaluation of gyandoot. 2003. http://www1.worldbank.org/publicsector/bnpp/Gyandoot.PDF
2. Mission 2007 project. 2005. http://www.mission2007.org
3. A. Chand. Designing for the indian rural population: Interaction design challenges. 2002.
4. R. Heeks. Information and communications technologies for development: A comparative analysis of impacts and costs from India. 2003.

5. D. R. Hutchings, G. Smith, B. Meyers, M. Czerwinski, and G. Robertson. Display space usage and window management operation comparisons between single monitor and multiple monitor users. In *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, pages 32–39, New York, NY, USA, 2004. ACM Press.
6. K. Kenniston. Grassroots ict projects in India: Some preliminary hypothesis. *ASCII Journal of Management*, 31(1 and 2), 2002.
7. S. Mitra. Self organising systems for mass computer literacy: Findings from the hole in the wall experiments. *International Journal of Development Issues*, 4(1):71–81, 2005.
8. D. Nitin Gachhayat. Personal correspondence. 2005.
9. B. Parthasarathy, A. Punathambekar, G. R. K. Guntuku, D. Kumar, J. Srinivasan, and R. Kumar. Information and communications technologies for development: A comparative analysis of impacts and costs from India. 2005. http://www.iiitb.ac.in/ICTforD/Complete report v2.pdf
10. R. Roman and C. Blattman. Research for telecenter development: Obstacles and opportunities. *The Journal of Development Communication.*, 12(2):745–770, December 2001.
11. K. Toyama, K. Kiri, D. Menon, J. Pal, S. Sethi, and J. Srinivasan. Pc kiosk trends in rural India. 2005. http://www.globaledevelopment.org/papers/PC Kiosk Trends in Rural India.doc