ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A key log Mining Technique to analyze web usage access pattern in an Organization for Internet access Security

K.Tarun Kumar,
Student, Dept. of CSE,
JNIT, India
poomanrun@gmail.com

N.Pradeep Kumar
Student, Dept. of CSE,
JNIT, India
Pradeepn530@gmail.com

A.ShilpaBhargavi
Student, Dept. of CSE,
JNIT, India
sai.shilpa43@gmail.com

Dr. G. Manjunath*
Professor and Head,Dept of CSE & IT,
JNIT, India.
gmanjunathc2000@yahoo.co.in

Abstract: One of the most comprehensive way of providing security to internet access is through firewall. Firewall is a mechanism which allows rule based internet access. Certain sites can be blocked, few sites can be allowed and few sites can be given restricted access through internet security. But internet is ever evolving. New sites come in everyday. It makes the firewall managers and Network administrator's Job very difficult to manage and restrict sites. In large corporate IP address based logging is enabled to view the sites the employees are accessing. One of the security policies includes logging of packets from which important information can be gathered regarding the type of sites or the contents accessed by the user. This technique has certain disadvantage in a sense that the information propagated are packet based and can only trace the files that are accessed through the internet infrastructure of the organization. There are several bypasses that can be designed to overcome this system for example the user may use an alternative internet access gateway like one through GPRS which entirely bypass the LAN logger or the firewall. This may also include data extraction through wireless interface like Bluetooth or wifi. Assuming the fact that personal data access is not allowed in such a corporate environment we propose a unique technique for accessing the internet activity of the user by logging the keystrokes and further extracting meaningful information from the logs. As user presses keyboard or mouse keys, it is logged by generating an interrupt to the kernel as a background process. The logged data is encrypted using RC4 cryptosystem with an administrative password. The log files are decrypted periodically and data is analyzed using data mining technique to get an overview of the activities of the user. The process is a background process and log files cannot be manipulated like that of internet access log. Hence the method is secured, efficient and well suited for pattern extraction from user internet access.

Keywords: component; formatting; style; styling; insert (Minimum 5 to 8 key words)

I. INTRODUCTION *(HEADING 1)*

A key logger, or keystroke logger, is a piece of hardware or software which records user keystrokes such as instant messages, e-mail, and any information we type at any time using your keyboard. Many key log solutions are very careful to be invisible to computer users and are often used by employers to ensure employees use work computers for business purposes only.

A. Simple Keystroke Mining Life Cycle

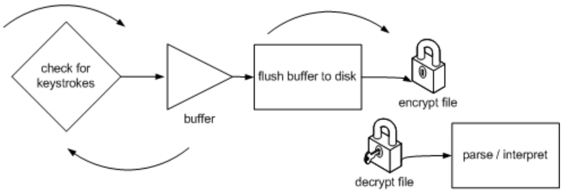


Figure 1. Simple Keystroke Mining Life cycle.

1. A key logger typically records actions and events on a computer to a volatile buffer.
2. Once the buffer is full, or on set intervals, the buffer is flushed to a non-volatile log file.
3. A common practice among key log solutions is to encrypt the log file.

The key logging mechanics illustrated are not stealth, and vastly inefficient. There are better, commercially available keyboard hooks and stealthy key loggers available.

B. Extracting the Pattern from Key mined File

Once details of pressed keys are extracted, the details are stored in a file. Now the most significant user activities must be obtained from this file. This is performed using TF-IDF measure with document analysis technique.

The term count in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) to give a measure of the importance of the term ti within the particular document d_j. Thus we have the term frequency, defined as follows.

$$\mathrm{tf}_{\mathrm{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is the number of occurrences of the considered term (t_i) in document d_j, and the denominator is the sum of number of occurrences of all terms in document d_j, that is, the size of the document $| d_j |$.

As the document or the log file contains various other keys like ALT/ENTER?PERIOD, it is first filtered to separate these keys and than the above technique is used to extract the most significant terms and sentences from the key logged pattern of the user.

Further the similarity analysis can be performed base don words or sentences.

Word Order Similarity:

The word order similarity is mainly used to describe the sequence similarity between two sentences. Chinese sentence can be presented by many kinds of style, the different sequence of the words stand for different meanings. Here we describe the sentence as three vectors as follows:

$V1=\{d_{11},d_{12},\ldots,d_{1n1}\}$

$V2=\{d_{21},d_{22},\ldots,d_{2n2}\}$

$V3=\{d_{31},d_{32},\ldots,d_{3n3}\}$

Here the weight d_{1i} in vector V_1 is the t_f-i_{df} value of the words; the weight d_{2i} in vector V_2 is the bi-gram whether occur in the sentence (0 stands for no-occurring, 1 stands for occurring); the weight d_{3i} in vector V_3 is the tri-gram whether occur in the sentence. The word order similarity between S_1 and S_2 is:

$$Sim2(S1,S2)=\lambda1*Cos(V11,V21)+\lambda2*Cos(V12,V22)+\lambda3*Cos(V13,V23) \qquad (2)$$

Here $\lambda1+\lambda2+\lambda3=1$. λi stands for the ratio of each part.

Word Semantic Similarity:

The word semantic similarity is mainly used to describe the semantic similarity between two sentences. Based on semantic similarity among words, we define word-Sentence Similarity (WSSim) to be the maximum similarity between the word w and words within the sentence S. Therefore, we estimate WSSim(w,S) with the following formula:

WSSim(w,S)=max{Sim(w, Wi)|Wi€S, where w and Wi are words} (3)

Here the Sim(w,Wi) is the word similarity between w and Wi. With WSSim(w,S), we define the sentence similarity as follows:

$$\mathrm{Sim}_3(\mathrm{S}_1,\mathrm{S}_2)=\frac{\sum_{w_i\in S_1} WSSim(w_i,S_2)+\sum_{w_j\in S_2} WSSim(w_j,S_1)}{|S_1|+|S_2|} \qquad (4)$$

Here S1, S2 are sentences; |S| is the number in the sentence S.

Sentence Similarity:

The sentence similarity usually described as a number between zero and one, zero stands for non-similar, one stands for total similar. The larger the number is, the more the sentences similar. The sentence similarity between S1 and S2 is defined as follows:

$$Sim(S1,S2)=\lambda1*Sim1(S1,S2)+\lambda2*Sim2(S1,S2)+\lambda3*Sim3(S1,S2) \qquad (5)$$

Here $\lambda1$, $\lambda2$, $\lambda3$ is the constant, and satisfied the equation:

$\lambda1+\lambda2+\lambda3=1$. In this paper, $\lambda1=0.2$, $\lambda2=0.1$, $\lambda3=0.7$. follow.

II. RELATED WORK

[1] presents a study on the ability and the benefits of using a keystroke dynamics authentication method for collaborative systems and focuses on biometric based solutions which donot require any additional

sensor. A serious threat to confidentaility is spyware which leads to loss of control over private data for users. Signature and Heuristic based detection is traditionally used and also performs well against

spyware[2]. [2] presentes a spyware detection approach based on Data Mining and also shown the method was successful. The legitimate user's typing patterns are combined with the user's password to

generate a hardened password whhich is convincingly more secure than conventional passwords against both online and offlineattackers and also proved that this approach is viable in practice and has ease of

use with improved security and performance[3].keystroke dynamics is a reliable security instrument for authentication, if combined with other instruments[4]. [6] presents a remote authentication framework

called TUBA for monitoring a user's typing patterns and evaluated robustness of TUBA. keystroke dynamics is robust against synthetic forgery attacks studied, where attacker draws statistical samples from a

pool of available keystroke datasets other than the target. TUBA is particularly suitable for detecting extrusion in organizations and protecting the integrity of hosts in collaborative environments, as well as

authentication[5]. A feature extraction methodology which is based on frequent sequence mining within and across multiple modalities of user input is applied for the fusion of physiological signals and gameplay

information in a game survey dataset. The obtained sequences are analysed and used as predictors of user affect resulting in computational models of equal or higher accuracy compared to the models built on

standard statistical features[6]. [7] proposes a new graphical-based password KDA system for touch screen handheld mobile devices. The graphical password enlarges the password space size and promotes the KDA utility in touch screen handheld mobile devices [7].

III. PROPOSED SYSTEM

There is various mining technique available and there are various key logging techniques are also available. The key logging is provided by windows services itself. But there is no comprehensive framework for a definitive key mining solution. The techniques are generalized and tracks only session wise data. Moreover the system needs to be manually started. The system proposed here gets started automatically as startup object. Once it starts, it keeps tracking the keys that are being hit by the user. Periodically this information is put back in the hard disk. Once the information is stored in a log file, administrator can extract this file and can fetch meaningful information out of them. The generated information is also stored in the secondary storage for later referrals.

The design fundamentals are presented as below.

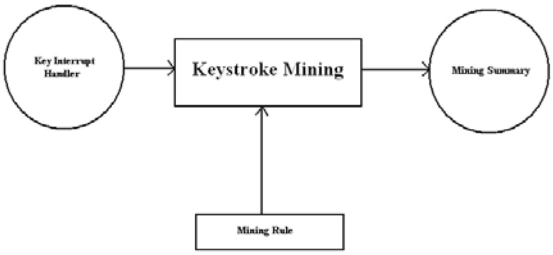


Figure 2: 1st level DFD

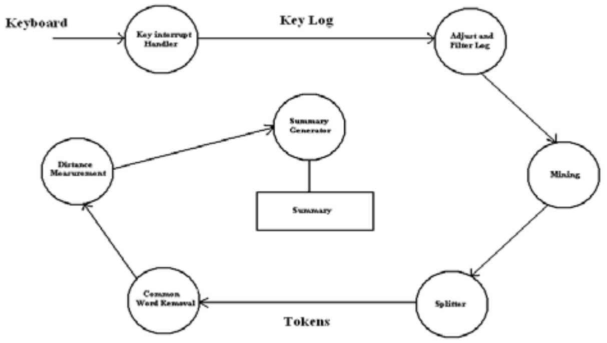


Figure 3: 2nd Level DFD

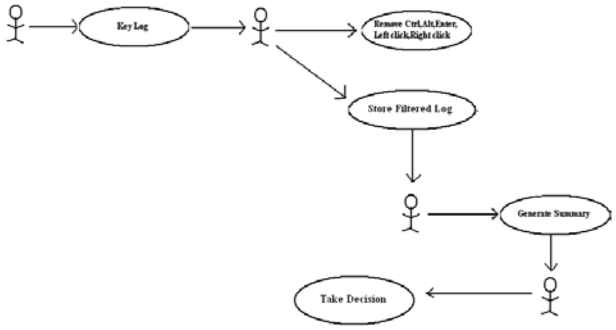


Figure 4: Use Case Diagram

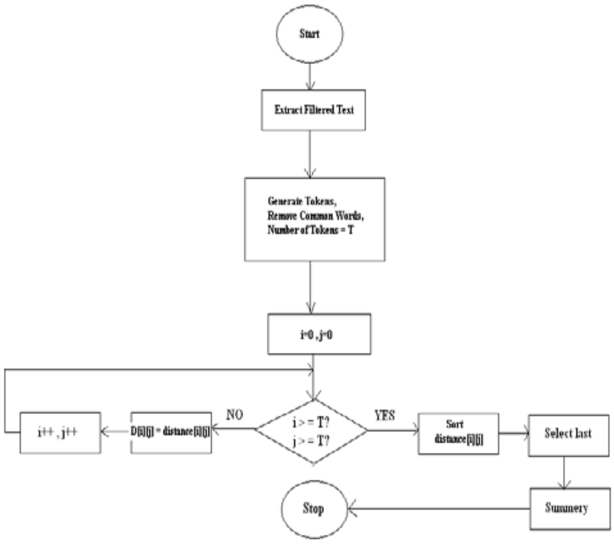


Figure 5: Flow Chart

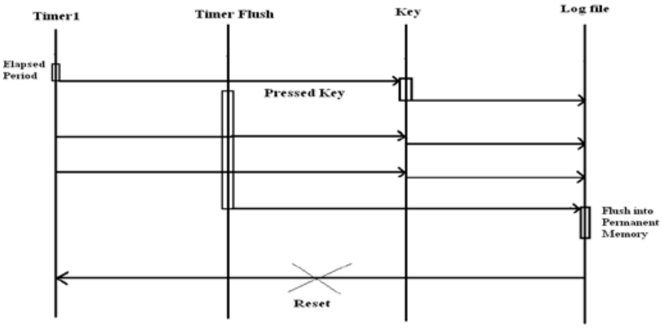


Figure 6: Sequence Diagram

IV. METHODOLOGY

1) Start a timer that ticks in every 10 milli seconds and looks for user key pressed

2) If any key is pressed, then the ascii code is extracted and stored in a local buffer

3) After every flush interval F, flush the buffer in a log file.

4) After every D interval, take a Snap shot of the desktop and store it in the log directory.

5) At every E interval where D>E>F, encrypt the log file

Mining Process

6) Partition the log files into sentences.

7) Remove the common words like the, am, is, are

8) If the key is delete or backspace than modify the contents

9) Extract Term frequency.

10) Select the highest term frequency as the summery of the log.

V. RESULTS AND ANALYSIS

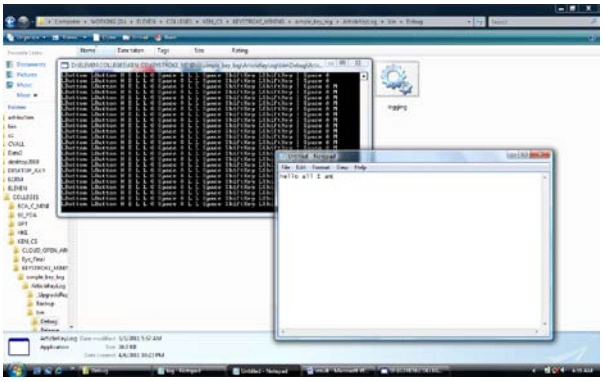
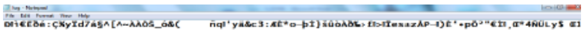


Figure 7. : Tracking of the keys



```
ÐÑ1€ÈÕé:ÇXyÏd7á§^[^-ÀÀÕŠ_ó&(      ñq!'yä&c3:ÆÊ*o-þÏ}šûòÀÕ‰›£Ð>ÏÏexazÀP-I)É'•pÕ²"€ÏI,Œ"4ÑÜLy$ Œ1
```

Figure 8. Encrypted Text

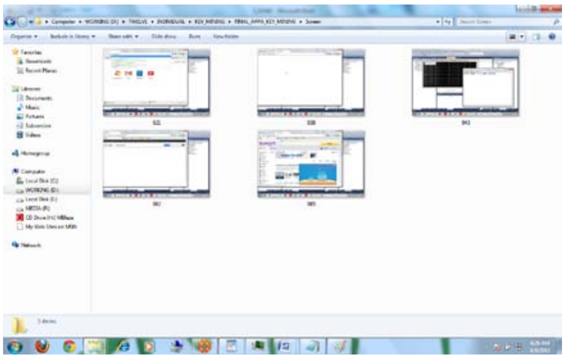


Figure 9. Background Snap of the Desktop

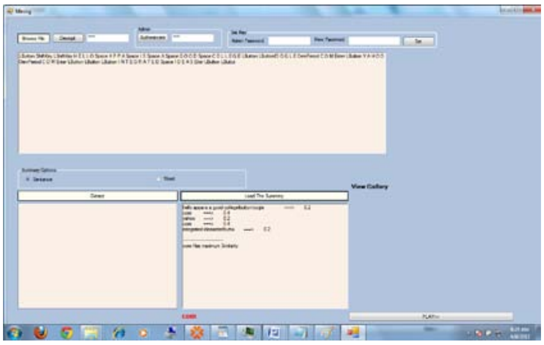


Figure 10. Main Interface and Summery Generation

Table I. Performance of the System

Total Logging time in Hours	Image Space(Mb)	Log Memory	Cpu Utilization	Mining Time(Seconds)
1	30	3Mb	3%	6
2	70	10Mb	3.80%	15
5	140	20Mb	3.10%	220
8	180	27Mb	4%	600

The performance of the system is extensively tested in a real time test scenario by keeping the process running in some of the busy university systems. It can be seen from the table that the CPU utilization does not increase with logging time which makes the system a stable one. The memory required on the disk by the log file and image files are proportional to the time of observation. But log file disk space depends upon the usage. Therefore it varies with time as par usage. But as the snap of the desktop is taken periodically, it is linear time dependent.

VI. CONCLUSION

Keystroke mining is a technique for extracting the user activities. Whenever the system is logged in the process starts automatically and starts monitoring the key that is being hit by the user. When the user uses back space, such corrections are also incorporated. Overall the system tracks the session wise user data. This tracking has many irrelevant data like the information about mouse click and so on. Such information can be filtered out and finally a comprehensive summery can be generated which can reveal useful information about the system access of the user. The project can be further improved by incorporating more stronger mining rule to link one session result with the other session. Further new adaptations can be made to encode the images to reduce the size of the images in the disk and data compression standard can be adopted for keeping the disk space of the log file minimum.

VII. REFERENCES

[1] R. Giot, M. El-Abed, and C. Rosenberger, “Keystroke dynamics authentication for collaborative systems,” in The IEEE International Symposium on Collaborative Technologies and Systems (CTS), 2009, pp. 172–179.

[2] R. K. Shahzad, S. I. Haider, and N. Lavesson, “Detection of Spyware by Mining Executable Files,” in Proceedings of the International Conference on Availability, Reliability, and Security (ARES 10), 2010, pp. 295-302.

[3] F. Monrose, M. K. Reiter, and S. Wetzel. Password hardening based on keystroke dynamics. In G. Tsudik, editor, Sixth ACM Conference on Computer and Communications Security, pages 73-82. ACM Press, 1999.

[4] S. Douhou and J. R. Magnus. The reliability of user authentication through keystroke dynamics. Statistica Neerlandica, 63(4):432–449, November 2009.

[5] D. Stefan and D. Yao. Keystroke-dynamics authentication against synthetic forgeries. In Proceedings of the International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), November 2010.

[6] Héctor P. Martínez and Georgios N. Yannakakis,"Mining Multimodal Sequential Patterns: A Case Study on Affect Detection", 13th International Conference on Multimodal Interaction, November 14-18, 2011.

[7] Ting-Yi Chang, Cheng-Jung Tsai, Jyun-Hao Lin, "A graphical-based password keystroke dynamic authentication system for touch screen handheld mobile devices", The Journal of Systems and Software, pp.1157-1165, 2012.