Survey and Analysis of the Event-log

Mr.Shirish Mohan Dubey, Mr. Rajeshwar Shukla and Dr. Udai Shanker

Abstract – In recent year, Event-log contain large amount of data. It is an important task for system management(Security Analysis) ,Network Administrator. Event-logs play an important role in modern IT systems. Since it is useful source for determining the present status of the system. Event Viewer maintains logs about program Security & System Events. Our intension is to use error analysis of event type. Data mining techniques are a common choice for knowledge discovery from Event-logs & the mining of patterns from Event-logs has been identified as an important System & Network management task. Recently proposed mining approaches for accomplishing this task have often been based on well known algorithm for mining frequent item sets. They have focused on detecting frequent event type patterns. The aim of this paper is to provide comparisons among different error type occurring in the system. We can formulate the task of mining frequent error event type patterns or frequent error item sets. We use Apriori Algorithm for finding frequent sets.

Keywords: **Data Mining, Event-log Analysis, Event Types, Logfile.**

I. Introduction

Data mining refers to extracting or “mining” knowledge from large amount of data. The information and Knowledge gained can be used for applications ranging from business management production control and market analysis to engineering design and science exploration. Data mining can be viewed as a result of natural evolution of information technology.

Event-logs are an important role for System management and network administrators. Logfiles are an important source for determining the present status of the system. An Event-log or Logfiles consists of several independent lines of text data, which contain information that refer to events into the system. The source of system health information No of tools have been developed for monitoring Logfiles. Swatch[z], logsurfer[3].The contents of Event-logs indicated the current status of the system(s) that they monitor. Logfiles techniques can be categorized into two category fault detection and anomaly detection.

In fault detection the domain expert creates a databases of fault massages patterns. If a line is appended to a Logfile that matches a pattern, the Logfile monitor takes a certain action. In anomaly detection, a system profile is created which reflects normal system activity. If messages are logged that do not fit the profile an alarm is raised. But creating the system profile by hand is time consuming and error prone. To solve two problems various methods have been proposed, with data mining methods being one of the most popular choices[5,6,7,8].In most research papers the focus has been on mining frequent patterns from Event-logs. This helps one to find patterns that characterize the normal behavior of the system and facilitates the creation of the system profile. Recent research papers have mainly proposed the mining of temporal patterns from Event-logs with various association rule Algorithms [5,6,7,8,9].These algorithms assume that the Event-log has been normalized i.e. all events in the Event-log have a common format. Association rule Algorithm have been often used for detecting temporal association between Event Type[5,6,7,8,9] e.g. if events of type C & D occur within 5 seconds they will be allowed by an event of E. Within 60 seconds (each detected temporal association has a certain frequency and confidence).Although Association rule algorithm are powerful they often can’t be directly applied to Logfiles because Logfiles do not have a common format. Logfiles lines rarely have all the attributes that are needed by the association rule algorithms. For example-the widely used syslog protocol does not impose strict requirements on the message format[1].A typical syslog message has just the timestamp , host name and program attribute that are followed by a free message string but only the message string part is mandatory[1].One important attribute that Logfile line often lack is the Event type. Fortunately it is possible to derive event types from log lines, since very often the events of the same type correspond to a certain line pattern .For example the Lines-

Router myrouter1 interface 192.168.13.1 down

Router myrouter2 interface 10.10.10.12 down

Router myrouter3 interface 192.168.22.5 down

Represents the Event type “interface down” and correspond to the line pattern. Router * interface * down. Line patterns could be identified by manually reviewing Logfiles, but this

Mr.Shirish Mohan Dubey and Mr. Rajeshwar Shukla are with MCA, GTU, University, Atmiya Institute of Technology and Science (AITS), RAJKOT (GUJRAT) INDIA, and Dr. Udai Shanker MMM Engineering College, Autonomous GORAKHPUR (U.P.), INDIA, [1]dubeys78@gmail.com , [2]iiitnrs@gmail.com , [3]udaigkp@gmail.com

is feasible for small Logfiles only. One charming choice for solving this problem is the data clustering algorithm. Clustering algorithm aim at dividing the set of objects(clusters) into groups where objects in each cluster are similar to each other and as dissimilar to objects from other clusters. Objects that are not fit any of the groups it is called outliers. When Logfile lines are viewed as objects, Clustering Algorithms are a natural choice, because line patterns form natural cluster-line that match a certain patterns are all similar to each other & generally dissimilar to lines that match other patterns. After the clusters (Event Types) have been identify. Association rule algorithms have can be applied for detecting temporal association between event types. Note that Logfile data clustering is not merely a preprocessing step. A Clustering Algorithm could identify many line patterns that reflect normal system activity & that can be immediately included in the System profile. Since the user does not wish to analyze them further with the Association Rate Algorithms. Moreover the cluster of outliers that is formed by the clustering algorithm contains infrequent lines that could represent previously unknown fault condition or other unexpected behavior of the System. In this paper we discuss existing data clustering and propose a new clustering algorithm for mining line patterns from Logfiles. The remainders of the paper is organized as follows:- Section II presents a overview of SLCT(Simple Logfile Clustering Tool); section III discusses related work on data clustering ;section IV presents clustering Algorithm Logfile data ; Section V describes the future work ;section VI conclude the paper.

II SLCT(Simple Logfile Clustering Tool)

SLCT employs a data clustering for analyzing textual Event-logs where log line represents a certain event. The data clustering problem can be defined as follows:- Divide a set of data points into groups so that points from the same cluster are similar to each other; point that do not fit well to any of the detected clusters are called Outliers. SLCT views each Event-log lines as a data point with categorical (Non Numeric) attributes ,where the Kth word of the line is the value of the Kth attribute. SLCT uses a density based method for clustering – if identifies dense regions in the data space and forms clusters from them.With each cluster corresponding to a certain frequently occurring line patterns. Since Outliers are able to detect dissimilar to clustered points, SLCT is also able to detect infrequent events that possibly represents serious anomalies in the behavior of the system.

III Related work on data clustering

Clustering method have been researched extensively over the past decades, many algorithms have been developed. The clustering problem is often defined as follows: given a set of points with n attributes in the $\mathbf{R}^n$ find a partition of points into clusters so that point within each cluster are close to each other. To determine, how close two points x and y are to each other, a distance function d(x,y) is employed .Many algorithms use a certain variant of Lp norm(p=1,2,3,--------) for the distance function.

$$\mathbf{d_p(x,y)= {}^{p}\sqrt{}\ \sum_{i=1}^{n} |x_i-y_i|^p}$$

Today, there are two major challenges for traditional clustering methods that were originally designed for clustering numerical data in low-dimensional spares(where usually n is well below 10)

IV Clustering Algorithm for Logfile Data

1) The nature of Logfile Data-

The nature of the data to be clustered plays a role when choosing the right algorithm for clustering. Most of the clustering algorithm have been designed for generic data sets such as market basket data. Where no specific assumption about the nature of the data are made. How ever when we inspect the content of typical Logfiles at the word level. There are two important properties that distinguish Logfile data from a generic data set.During our experiment these properties are relevant. We used six Logfile data from various domains: HP Open View Event-logfile,Mail Server Logfile(The Server are running sendmail, ipopd & imapddaemons), Squid Cache Server Logfile , File & Print Server Logfile & Win 2000 domain controller Logfile.It is impossible to verify that the properties we have discovered characterize every Logfile ever created on earth, we still believe that they are common to a wide range of Logfile data set.We analysis only event error log data. In Windows –XP an event is significantly occur in the system or in a program that requires users to be notified or an entry added to a log. With the Event-logs in Event Viewer, you can obtain information about your hardware and software & system components and monitor security events on a local or remote computer.

V Future Work and Availability Information

For a Future work we plan to investigate various association rule algorithm, in order to create a set of tools for building Logfile profiles. We will be focusing on algorithm for detecting temporal patterns, but also an algorithms for detecting association between event attributes within a single event cluster.

VI Work done and Analysis

We tried to give detail information to industry that worked or operating system to remove errors so we do firstly thorough study on event viewer of window operating systems(Client Server) .Do analysis on every error which is occurred in run time or at the time of installation of any application. also analysis the type of error which is frequently occurred when operating system is running. We also save the Logfiles from different type of Operating System in different format for analysis & making approach for appropriate result and graph. We also do research & Development to apply mining rules & algorithm in this data for extracting useful information only error type.

VII Conclusion

Our analysis gives priority wise report about the errors. It helps Vendor to get information about error and fixing bug related issues. It also helps to make software more maintainable & scalable. Users can get the quick solution from the vendor. Our analysis is approached to give useful information for relevant industry in the form of graph or report which is related to time and particular errors.

APPENDIX

A Sample Centralized Logging Infrastructure:

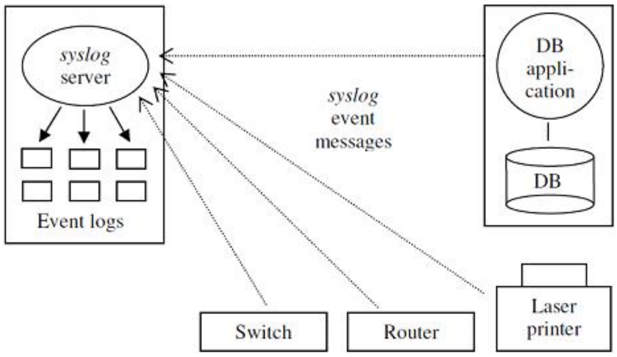

Fig: A Sample Centralized Logging Infrastructure

Applications, servers, and network devices use the *syslog* protocol for logging their events to the central log server that runs a *syslog* server. Log monitoring takes place on the central log server and alerts are sent to the monitoring console.

- Event – a change in the system state, e.g., a disk failure; when a system component (application, network device, etc.) encounters an event, it could emit an event message that describes the event.
- Event logging – a procedure of storing event messages to a local or remote (usually flat-file) event log.
- Event Type- The description of each event that is logged depends on the type of event. Each event in a log can be classified into one of the following types

The description of each event that is logged depends on the type of event. Each event in a log can be classified into one of the following types:

• *Information*

An event that describes the successful operation of a task, such as an application, driver, or service. For example, an Information event is logged when a network driver loads successfully.

• *Warning*

An event that is not necessarily significant, however, may indicate the possible occurrence of a future problem. For example, a Warning message is logged when disk space starts to run low.

• *Error*

An event that describes a significant problem, such as the failure of a critical task. Error events may involve data loss or loss of functionality. For example, an Error event is logged if a service fails to load during startup.

• Success Audit (Security log) an event that describes the successful completion of an audited security event. For example, a Success Audit event is logged when a user logs on to the computer.

• Failure Audit (Security log) an event that describes an audited security event that did not complete successfully. For example, a Failure Audit may be logged when a user cannot access a network drive.

VIEWING EVENT DETAILS:

The following is an example of an event that does not show information properly.

Event Type: Information
Event Source: MSExchangeIS Private
Event Category: (30)
Event ID: 2003
Date: 8/16/2001
Time: 1:47:02 PM
User: N/A
Computer: SERVERNAME
Description: The description for Event ID (2003) in Source (MSExchangeIS Private) cannot be found. The local computer may not have the necessary registry information or message DLL files to display messages from a remote computer. This example is the event as it appeared when opened on a computer without Exchange Server.

VIII References

[1] C. Lonvick, “The BSD syslog Protocol”, *RFC3164*, 2001.

[2] Stephen E. Hansen and E. Todd Atkins, “Automated System Monitoring and Notification With Swatch”, *Proceedings of the USENIX 7th System Administration Conference*, 1993.

[3] Wolfgang Ley and Uwe Ellerman, logsurfer(1) manual page, unpublished
(see http://www.cert.dfn.de/eng/logsurf/), 1995.

[4] Risto Vaarandi, "SEC - a Lightweight Event Correlation Tool", *Proceedings of the 2nd IEEE Workshop on IP Operations and Management*, 2002.
5] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of frequent episodes in event sequences", *Data Mining and Knowledge Discovery*,Vol. 1(3), 1997.

[6] M. Klemettinen, H. Mannila, and H. Toivonen, "Rule Discovery in Telecommunication Alarm Data", *Journal of Network and Systems Management*, Vol. 7(4), 1999.

[7] Qingguo Zheng, Ke Xu, Weifeng Lv, and Shilong Ma, "Intelligent Search of Correlated Alarms from Database Containing Noise Data",*Proceedings of the 8th IEEE/IFIP Network Operations and Management*
Symposium, 2002.

[8] L. Burns, J. L. Hellerstein, S. Ma, C. S. Perng, D. A. Rabenhorst, and D. Taylor, "A Systematic Approach to Discovering Correlation Rules For Event Management", *Proceedings of the 7th IFIP/IEEE International Symposium on Integrated Network Management*, 2001.

[9] Sheng Ma and Joseph L. Hellerstein, "Mining Partially Periodic Event Patterns with Unknown Periods", *Proceedings of the 16th International Conference on Data Engineering*, 2000.

[10] Matt Bing and Carl Erickson, "Extending UNIX System Logging with SHARP", *Proceedings of the USENIX 14th System Administration Conference*, 2000

Shirish Mohan Dubey received His MCA degree from Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, M.P, India in year 2003. He was received M.Tech (CSE) degree from Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, M.P, India in year 2011. Currently he is working as an Assistant Professor (MCA) in Atmiya Institute of Technology and Science (AITS), Yogidham Gurukul, Kalawad Road, Rajkot, Gujarat, affiliated to Gujarat Technological University (GTU) Ahmedabad. He has been teaching various Computer Science/Information Technology subjects such as Data Mining and Warehousing, Operating system and Mobile Computing, Data structure since 2003. His area of Research interest includes effective algorithm for Data Mining and Warehousing. E-mail: dubeys78@gmail.com

Rajeshwar Shukla received his B.Sc degree from Deen Dayal Upadhyay Gorakhpur University, Gorakhpur, U.P, India in year 2000 and received MCA degree from Maharshi Dayanand University, Rohtak, Haryana, India with First Class in Jan 2008. He was received M.Tech (IT) degree from Karnataka State Open University, Mysore with Distinction Class in Jul 2010. He is Pursuing PhD(CSE)* from Shri Jagdishprasad Jhabarmal Tibrewala University, Vidya Nagari, Jhunjhunu-Churu Road, Chudela, Dist. Jhunjhunu, Rajasthan in year 2012. Currently he is working as an Assistant Professor (MCA) in Atmiya Institute of Technology and Science (AITS), Yogidham Gurukul, Kalawad Road, Rajkot, Gujarat, affiliated to Gujarat Technological University (GTU) Ahmedabad. He has written Nine (9) Computer Science/Information Technology Books for Pragya Publications Pvt. Ltd. E-38, Industrial area, Mathura, (U.P). He has been teaching various Computer Science/Information Technology subjects such as C, C++, VB, .Net, DBMS, Unix\Linux, Mobile Computing, Multimedia and web technology since 2003. His area of Research interest includes Web based education/E-Learning, DBMS, data warehousing and data mining and Grid Databases. E-mail: iiitnrs@gmail.com, rvsaits@gmail.com

Dr. Udai Shanker received His BE (Electrical Engineering) degree from M.M. M. Engg. College, Gorakhpur, U.P. India in year 1986 with First (71.6%) Class and received M.E. (Computer Engineering) degree from Department of Electronics & Telecommunication Engineering, Jadavpur University, Calcutta, India with First (76.4%) Class in March 1998. He received PhD (Computer Engineering) from Department of Electronics & Computer Engineering, Indian Institute of Technology Roorkee, Roorkee-247 667, India in June 2006. His Topic of the Thesis was Some Performance Issues in Distributed Real Time Database Systems. Currently he is working as a Professor & Head Department of Computer Sc. & Engineering, M. M. M. Engineering College, Gorakhpur-273 010, Uttar Pradesh, India. He has published several research papers in national/ International Journals and National/International Conferences. He has also written many Book Chapters for various National/International publishers. His area of Research interest includes Real Time Systems, Distributed Real Time Database Systems, Mobile Distributed Real Time Database Systems and Grid Databases. He is Member of many Professional Societies such as Institution of Engineers (India) *(Life Member-MIE-M061087-3)*, Indian Society for Technical Education (*Life Member-LM-14307*), Computer Society of India (*Life Member-10464*), The Institution of Electronics and Telecommunication Engineers (*Fellow, F-212107*). E-mail: udaigkp@gmail.com, udaigkp@rediffmail.com