Research Article

CONVERSION LOG DATASET INTO UNDERSTANDABLE FORMAT FOR DATA MINING

*Ali Azimi Kashani, Alimohamad Monjezi Noori and Hadi Mahriyar

Department of Computer Engineering, Shoushtar Branch, Islamic Azad University, Shoushtar, Iran *Author for Correspondence

ABSTRACT

In recent years, the importance of log analysis has grown. Log data constitute a relevant aspect in understanding user behaviour. The volume of log is increasing rapidly and the format is differ to each organization. Having knowledge in analysing log would be an important part in log analysis. An efficient data pre-processing technique can help to improve the quality of data thus help to improve the accuracy and the efficiency of the subsequent data mining analysis. In this research, systematic analysis is done on existing techniques to transform raw dataset into readable format. Hence, transforming the webserver logs into a database for data mining analysis with the aims to help researchers gather knowledge about log data.

Keywords: Log Analysis, Data pre-processing, Data Transformation, Data Mining, Knowledge Database Discovery, Systematic Analysis

INTRODUCTION

Nowadays, the importance of log analysis has grown. Log data constitute a relevant aspect in understanding user behaviour. When user accesses to the website, log files are created. Log file recorded all related information regarding to the user such as host name, IP address, date and time, URL and request type. Information are usually stored in log file by using different log format such as NCSA (National Center for Supercomputing Application) common log, W3C Extended Log File and IIS log file format. The log file can be found on different location, for example, in web server, web proxy server and client browser (Patil, 2012). In order to understand the results of search engines or user behaviour, applications and computer system, it is necessary to do analysis on log file related to the website. However, the volumes of log nowadays have growing rapidly in respond to the tremendous uses of web. Therefore, there is a crucial need for a new generation of theories and tools to assist human in extracting useful information.

Knowledge database discovery (KDD) is considerable importance and necessity because of the accessibility and abundance of data today (Maimon and Rokach, 2005). KDD processes comprise of nine steps which are domain understanding and KDD goals, selection and addition, pre-processing, data transformation, choosing the appropriate data mining task, choosing the data mining algorithm, employing the data mining algorithm, evaluation and utilize the discovered knowledge (Fayyad *et al.*, 1996). For this research, the scope of dataset will be the USIM web server log and KDD technique was used to do the data transformation.

Data transformation is the process where the data are transformed or consolidated into forms appropriate for subsequent analysis. The data are presented in certain representation in order to be transformed into nominal data. This research aims to transform the raw dataset into understandable format for subsequent data mining analysis. The domain categorisation were conducted and obtained prior to the data transformation process.

Apart from that, the motivations for this research are:

i. The Importance of how to Analyse Data and Transform Data into Readable Format

There are many information display in the raw log data. Analysing web server log helps to understand who visited the website, on what application, how often and also provide with information on errors, problems and performance problems of the web application. However, analysing the data without having knowledge to analyse the log would lead to wrong interpretation.



Research Article

ii. To Analyse log would be a Daunting Task

Nowadays, the size of log is increasing. The level of difficulty to analyse log become higher proportional to the number of user's accessess.

iii. The Difficulty for the Researcher to Transform Raw Dataset into Readable and Understandable Format for Subsequent Analysis

Transforming raw dataset into readable and understandable format need to be carried out prior to data mining analysis in order to ensure the outputs produced in the subsequent analysis will have a better accuracy and low false positive rates.

This research paper aims to transform the raw dataset based on the motivation above. The log dataset was obtained from the Pusat Teknologi Maklumat (PTM), Universiti Sains Islam Malaysia (USIM).

This paper is organised as follows. Section II presents the related works with big data and log. Section III explains the methodology used in this paper which consists of general overview on process to transform log dataset into understandable format and lab environment. Section IV presents the research finding which consists of log dataset classification, data transformation and machine learning algorithm results and section V is a closing remarks and summaries the future work of this research paper.

Related Work

Big Data

The variety, velocity and volume of data coming into an organization continues to reach unprecedented levels. This phenomenal growth of data requires not only to understand the big data but also what do with it. Table 1 shows the summarisation of the challenges of handling big data.

Author and Title	Challenges
Michael and Miller (2013) Big Data: New Opportunities and New Challenges	New challenges with respect to how much data to store, whether the data will be secure, how long it must be maintained and how much it will cost.
Hemerly (2013) Public Policy Considerations for Data-Driven Innovation	Policy makers must take into account the possibility that regulation could preclude economic and social benefits in order to achieve the maximum benefits from data-driven innovation.
Tallon (2013) Corporate Governance of Big Data: Perspectives on Value, Risk and Cost	Developing governance mechanisms, policies and structures become the challenges to the organisations. It strike the balance between reward and risk in the face of growing quantities of data and innovation that deliver cheaper storage technology, faster and better.
Pitt <i>et al.</i> , (2013) Transforming Big Data into Collective Awareness	Numerous challenges revolving around the notion of justice including distributive justice, procedural justice, interactional justice, natural justice and retributive justice. Another challenges include to resist the spread of misinformation.
Wagon and Clarke (2013) Big Data's Big Unintended Consequences	Data quality, semantic coherence and legality appear to be of little concern to those responsible for national security applications.

Table 1: Summarisation of the challenges of handling big data



Research Article

Log

Log file is an auxiliary text files that software application often produce (Valdman, 2011). Logging is a process of recording events or statistics to provide information about performance or system use (Bishop *et al.*, 1996). Log is used for record data on what, when, who, where and why (W5 questions) an event occurred for security professional on a particular application or device. Log is composed of log entries where information related to a specific event that has occurred within a network or system are contain in each entry (Allen, 2001). Different web servers maintain different types of information in the log files (Ratnesh *et al.*, 2009). The basic information in the log files are;

[x,y,z,a,b,c,d,e,f]

• User Name (x): Identifies who had visited the website. The identification of user is based on the IP address that is assigned by the Internet Service Provider (ISP).

• Visiting Path (y): The path taken by the user while visiting the website by using URL directly or by clicking on the link or through the search engine.

- Path Traversed (z): Identifies the path taken by the user within the website using several of links.
- Time Stamp (a): The time or session spent by the user in web page while browsing.
- Page Last Visited (b): The last page visited by the user before he or she leaves the website.

• Success Rate (c): The success rate of the website which determined by the number of downloads made and the number of copying activity done by the user.

- User Agent (d): User's browser information.
- URL (e): The resource that had been accessed by the user.
- **Request Type (f):** The information transfer method such as GET and POST.

The activity is recorded in web log file when user submit request to a web server.

Nowadays logs serve many functions within an organizations such as optimizing network and system compared to previous era which used primarily for troubleshooting the problems (Stout and Kent, 2002). Log is useful when involving with audit and forensics analysis, establishing baselines, supporting internal investigations and identifying operational trends and long-term problems (Ahmed *et al.*, 2009). Log file is used to study a user's query behaviour while user navigates a search site (Saxena *et al.*, 2012). Understanding user's navigational preferences is the way to improve query behaviour. The user access patterns information help service provides to modify and adapt their sites interface for individual users as well as to improve the site's static structure within the wider hypertext system.

Log files can be complex and often very large. Analysing log files could be a difficult task although the process of generating log files is straightforward and simple.

MATERIALS AND METHODS

Methodology

The followings are the procedure to transform log dataset into understandable format.



Research Article



Figure 1: General overview on process to transform log dataset into understandable format *Dataset*

The first step was cleaned up the dataset. The dataset is cleaned in excel format. Figure 2 shows the log format.

	wallarie	IT FORMALINE DO	to shree yes						(high)
N.C. Raberto Name Room	11444 - 11 11 11 11 11 11	6 x = = 0 ↓ A + = =	· · · · · · · · · · · · · · · · · · ·	a closer - S - Sc +	- (B)			Rine Arts	*
(taking)	to Inte		manuel	4 7444		Base	100	1.0 in	
	A	in months							
. 04					1.0				
							Frank 185	which a Research of	2
							seects we	eoute browsing L	op
(per l Name	Japoch Welterin Bro	99 PK (SK)							
arrenter y	1044518 reports of	while og the constituent							
								Phar Candilles	
mp have	- H							Section 1995	
shale brighting	1184								
chill hitting	1015-04-08 - 2015-	64-20							
chalt brittens Ex	1922-94-06 - 325-9- 1822-94-06 - 325-9-	Q#-32							
chide britgen g Ex Star Zian	18:25-34-08 - 302-3- 18:25-34-08 - 302-3- 18:26-30 - 22-49-34 28:1-40 Margodial	04-88							
chilt brighting Chi Ban Chill Chill Mont By	1005-04-00 - 303-0- UD-00-02 - 22-09-04 Decired Material 46	04-85							
etable belgenog die bee stim herd By spect Blade	1911 1912-94-09 - 3018- UD-80-00 - 20-Ph (H) Demol Mu (colum) 46 Treater	04-80						-1521/121	
chult Suligroop die dee dee tool By open I Blade	The Introduction and the Internet Managerial Aliantian Internet Managerial	04-23				121	3.0	(maintenant	
full Solgroup for en Sole sol By per Litter	Incl. Hold State	04-10 Jonus Name	Kool JP	ühdedine forminden	sheet By:	Tees	-	Tender (M. Romald	lares
fulkt telligting ne en tilse and thy gent tillede	Inc. 16(2)-16(0) - 2(2)-1 Inc. 16(2)-16(0) - 2(2)-16(16) Inc. 16(16)-12(2)-16(16) Inc. 16(16)-12(16)	SHEEL	Read JP concises [12]	Website Breaking	Marci By	Tens Intel-OF-DE DE TREAS	Ninkalay.	State Secure	Larry
fullt tellgring ny en tion ant fly ant tillede	Incl-Berg, 2015 Incl-Berg, 2015 Development Provide Automatic Atomic Income Incl Income Incl Income Incl Income Incl Income Incl Income Incl Income Incl Income Incl Income Income Income Income Income Income Income Incom	SHID	Road # 400 2002112 102 2002112	United to Basedag printersections printersection	Marcel By PELLIPSEY DOLUMENTY	Res 00109000000000000000000000000000000000	aliadada e	Inicial Anual Distances	dam
fulle Turligence re en ten ten ten ten ten ten ten ten ten	1905-04-00 - 2016 1905-04-02 - 20-05 1906-02 - 20-05 1906-03 1	See 25	Real # 00.3457512 05.0617612 05.0017612	Wednell + Desconding printersound com printersound com	Manet By D'ALLENDY althouty althouty	Test 1010-04-00 (1010-04 1010-04-00 (1010-04 1010-04-00 (1010-04 1010-04-00 (1010-04	Noteste	Interio Small Disa Second	dam
fade Seigneg re en tim and By and By en 1 Blade	1015-04-05-3035- 1015-04-05-3035- 1016000-20-4034 2016000-20-4034 40 1016000- 1016000- 101600-	Seatt	Read JP 00.2007b12 00.5007b12 00.5007b12 00.5007b12	Website & soundage purchased ages and mission datase purchased ages purchased ages	Marriel By PELLIGRAPY SOLUTIONARY SOLUTIONARY SOLUTIONARY	Res 00.0 (44.0) (10.0016) (0.0) (44.0) (10.0016) (10.0) (40.0) (10.0016) (10.0) (40.0) (10.0016)	Niversite .	Bini (Keuli Disebarrat	dam
Auto Sufficience No. Her Topo and By per 1 Harte No.	1012-04-00-2016 1012-04-00-2016 1010000-20-05-04 40. 1010000-2016 1010000-2016 1010000-2016 1010000-2017 101000-2017 101000-2017 101000-2017 101000-2017 101000-2017 101000-2017 101000-2017 101000-2017 10100-2017 10000-2017 1000-2017 10000-2017 1000-200	0.000 densign Massee (2000A.W.B) (2000A.W.B) (2000A.W.B) (2000A.W.B) (2000A.W.B) (2000A.W.B)	Rad # 0124347512 145047512 145047512 043047512 043047512 043047512	Vehicle Drawing printrocoll con printrocoll con printrocoll con printrocoll con printrocoll con printrocoll cont printrocoll cont printrocoll cont printrocoll cont printrocoll cont printrocoll cont printrocoll cont characterization	Manet By 1991-1992 1991-1992 1991-1992 1991-1992 1991-1992 1991-1992	Term 2012-04-00 07 20 70 00 2012-04-05 20 70 59 2012-05-05 00 20 20 10 2012-05-05 00 20 20 10 2012-05-05 00 20 20 10 2012-05-05 00 20 20 10	Ninkali e	Statute Recall	lane
Add Shilpinoi W M M M M M M M M M M M M M M M M M M	10(1-0+00-2014) 10(1-0+00-20-004) 10(1-0+004) 10(1-0+004) 10(1-0+04) 10(1	dead	Read # 00230025122 0050025122 0050025122 0050025122 0050025122 0050025122	Website Disserting professional can professional can professional can professional can professional can professional can	Marci In Philipping Philipping Philipping Philipping Philipping Philipping	Test 10.0 0440 (2000) 10.0 040 (2000) 10.0 040 (2000) 10.0 040 (2000) 10.0 040 (2000) 10.0 040 (2000)	Notesta -	Disco Rewall	Lines
Self Subjected W and State Anno 1 Theole Anno 1 Theole	101 301 301 101 301 301 301 101 301 301 301 301 101 301	Annue Namer (1994) (199	Real # 022402312 033002312 033002312 033002312 033002312 033002312 033002312 033002312 033002312	Webder Brescher gestassendigen gestassendigen gestassendigen gestassendigen gestassendigen unterstehendigen gestassendigen dieses zum	Material Ry Distantionary and colorary and colorary and colorary and colorary and colorary	Test 1012-0+62-227036 102-0+62-227036 102-0+62-227036 102-0+62-227036 102-0+62-227036 102-0+62-22703 102-0+62-22703 102-0+62-2555 102-0+62-255 102-0+62-55 102-0+62-555 102-0+62-555	Website	Status Status Status Status	dam
leidt Sufigened W er Stan Ver Ry er i Stedt	94 1922-94-00, 203-9 204-95 93 204-95 46 204-95 93 204-95 46 204-95 2000	Anong Name (association) (asso	Read JP 002 AbstPh 12 105 constitution 100 constitution 1	Weisleit in Sensolug performance and a sensolution generation and a sensolution and a standard as an and a standard and performance and a sensolution in the resolution of the sensoluti	Manual By PRAJECTORY DECLOSING DECLOSING PREJECTORY PREJECTORY PREJECTORY PREJECTORY PREJECTORY	Tase Dist 4940 (12703) Dist 4400 (2275) Dist 4400 (2275) Dist 4400 (2275) Dist 4400 (2275) Dist 4400 (2255) Dist	Nutrito	Inicip (pull	Jacob
Subit Subjections Server Taxa and By and Bhade	944 1922-34-00, 203-5 2025-02 - 204-5 2025-02 - 204-5 2025-020-020-5 2025-020-020-5 2025-020-020-5 2025-020-020-020-020-020-020-020-020-020	Annual Name Annual	500 8 500 30175 (2 500 30175 (2 500 30175 (2 500 30175 (2 500 30175 (2 500 30175 (2 500 30175 (2)	Website & meshap printication and any printication and printication and printication and printication of the state of the state of the printication and printication and printication and printication and printication and printication and printication and printication and printication and printication and printication and printication and printication and principation and principation and principation and principation and princip	Material Res Del La Decrary Selia La Decrary	Tent 10.1 (1993) 2.1 (1993) 10.1 (1993) 2.1 (1993) 10.2 (1993) 2.1 (1993) 10.1 (1993) 2.1 (1993) 10.2 (1993) 2.1 (1993) 2.1 (1993) 10.2 (1993) 2.1 (1993) 2.1 (1993) 10.2 (1993) 2.1 (1993) 2.1 (1993) 2.1 (1993) 10.2 (1993) 2.1 (19	Webster	Status Recall Bitte based	Asses
rhade Schlipping Mi Her Family Ann By Ann Bhy Ann Bhy A	94 1923-94-00, 203-94 2023-94-00, 203-94 2049-	Along Many Constant States (2004 With) (2004 With) (20	Read # 000402712 000402712 000407131 000407131 000407131 000407131 000407131 000407131	Website Tomonogy professional com- generational com- generational com- sistential complexity of the professional com- plexity of the profession of the source of the profession of the profession of the profession of the profession of the profession	Manual By DPA 10 Party PAL 10 Party	Tase 10.0 +0.0 (2700) 10.0 +0.0 (2705) 10.0 +0.0 (2005) 10.0 +0.0 (2705) 10.0 +	Nindradi tr	Inicipational Discharged	Aares
shelt bilgenig for en en en en en for on for on for en en for en en en en en en en en en en en en en	94 1023-9400-2014 2020-02-2014 2020-02-2014 40 2020-02-2014 2020-02-200-0200-0200-0200-0200-0200	Annual Manuel Science Manuel Science With Science With Sc	500 87 500 2017 12 500 2017 12	Before Descology professional con- generation sector con- analysis sector con- analysis sector con- trainers of the sector con- professional for sector con- professional for sector con- professional sector con- professional sector con- professional sector con- mention periodic sector con- tension and account of the sec- tor con- sector contrainers of the sector con- tension periodic sector con- contension periodic sector con- tension periodic sector con- tension periodic sector con- tension periodic sector con- tension periodic sector con- contension periodicon periodic sector con- contension periodic sector con- conten	Material Rey Profit Life (1999) Profit (1999) Profit Life (1999	Tens 100 (1994) (100 (1994) 100 (1994) (100 (1994)) 100 (1994) (100 (1994)) 1	Webster	Inte in forunt Alter bornet	Jamo (
chait highnig Mr Chai Chai Chai Chai Chai Chai Chai Chai	94 94 10223400 20.000 202000 20.000 202000 20.000 202000 20.000 202000 20.000 202000 20.000 202000 20.000 202000 20.000 202000 20.000 2020000 20.000 20200000 20.000 2020000000 20.000 2020000000000 20.000 2020000000000000000000000000000000000	development development (conserve with) conserve with conserve	Real # 00.0427512	Website & necessary grindroscol, con grindroscol, con gri	Manual Bay Ministry	Team 1012 0+14,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000 1012 0+04,0 21,0000	Nindora de	Shirp: Smull	Aares

Figure 2: Log file in xlsx format

Based on the log, the data are separated into 15 tabs which are user, group names, host IP, website browsing, identify by, time, website, attachment, access control policy, website link, destination IP, service port, application type, URL type and action. However, in this research only 3 tabs will be extracted along with its data which are website link, service port and URL type. The other information is not important as it may have same data, dynamic host IP, or incomplete data. The dataset will then be classified into three categories which are website link, service port and URL type.

© Copyright 2014 / Centre for Info Bio Technology (CIBTech)

Research Article

Research Environment

The laboratory is set up with one pc only. The lab use virtual private network for domain classification. Figure 3 shows the architecture of the laboratory.



Figure 3: PC 1 for storing and log files analysis

The following in Table 2 is the lists of the tools used in this lab. Almost all tools used are an open source software or free basis except Microsoft Excel.

Table 2: Tools list				
Function	Software/Hard	ware		Description
Display log file	Excel			Display log dataset in xlsx format.
Data mining tool	WEKA			Cluster and classify the new format dataset.
Storing and testing tool	PC 1			For storing log dataset and testing the dataset in WEKA.
Domain	Check	URL	category	Identifying the web URL type for
classification	(http://www.com	nmtouch.com/	url-miscat/)	Others category. It is an open source tool.

RESULTS AND DISCUSSION

Findings

The following are the findings of this research. A new dataset classification is produced to ease the data transformation process. The output of the new clean dataset is transformed into nominal dataset which can be used for the subsequent analysis using data mining algorithm (Mertz and Murphy,1996). WEKA software is used to conduct the mining analysis.

Log Dataset Classification

The log dataset has been classified based on three classification which are website link, service port and URL type as displayed in Figure 4.



Figure 4: Log dataset classification format

Research Article

Data Transformation

Dataset 1			New format for	
Website link	Service port	URL type	dataset 1	
nangastream.com/ 80 Entertainment		P11, P22, P37		
	5	ж 11	a collector	
the second s				
og dataset attributes		Datase	et transformed into nominal dat ic	
.og dataset attributes		Datase	et transformed into nominal dat it	
og dataset attributes 11 represents websit	e link – as mangasts	Datase numer ream.com/	et transformed into nominal det in	
og dataset attoihutes 111 represents websit 122 represents servico	e link – as mangasts 1 port – as port 80	Detroe numer ream.com/	et transformed into nominal dat ic	

Figure 5: Data Transformation (using certain number representation)



Figure 6: arff format of dataset



Figure 7: Arff file was uploaded in WEKA machine

The researcher identified the nominal data for each of dataset. Figure 5 shows the data transformation details. After conducting data transformation for each of dataset, the researcher came out with clean log dataset file that was compatible with WEKA machine. Figure 6 show the new format of dataset.

The new file is an arff format. The arff file will be used as input in WEKA machine to validate. Figure 7 show the arff file was uploaded into WEKA machine.

The purpose of this figure is to show how the data can be mined in WEKA software. Later, the data can be analysed using different machine learning algorithm inside WEKA.

Machine Learning Algorithm Results

This section presents the finding results of the dataset built. Naïve Bayes and k-nearest neighbour (IBk) algorithm is chose to classify the dataset and the result as displays in table 3. Both results from two

Research Article

algorithms were compared. True positive rate (TPR) and false positive rate (FPR) are used during the experiment. The experiment was conducted using WEKA software.

	Naïve Bayes (%)	IBk (%)	
TPR	100	99.6568	
FPR	0	0.3432	

Table 3: Machine learning algorithm result

TPR represents true positive rate, FPR represent false positive rate

Table 3 shows the result of TPR for Naïve Bayes algorithm is 0.3432% higher than IBk algorithm which indicates a good result.

Conclusion

As a conclusion, this research manage to provide a new clean log dataset to be published to help researchers in data mining research. Besides that, the dataset was presented in nominal data which was compatible to be used directly in WEKA machine leraning algorithm for data mining process. Furthermore, based on the experiment conducted using WEKA software, the True Positive Rate (TPR) of the classified data 100 % is produced. This result can be used as reference and comparison by other researchers with the same interests.

For future work, other bigger and different format of dataset is used. Other than that, different classifier algorithm will be used to be compare with.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to Universiti Sains Islam Malaysia (USIM) for the support and facilities provided. This research paper is supported by Universiti Sains Islam Malaysia (USIM) grants and FRGS grant: [PPP/FST/SKTS/30/12712],[PPP/FST/SKTS/30/12812] and [FRGS/1/2014/I CT04/USIM/02 /1].

REFERENCES

Ahmed MKMH and Raza A (2009). An Automated User Transparent Approach to log Web URLs for Forensic Analysis. *International Conference on IT Security Incident Management and IT Forensics*, Fifth 2(1).

Allen S (2001). Importance of Understanding Logs from an Information Security Standpoint. *GSEC* v1.2f. SANS Institute.

Bishop M, Wee C and Frank J (1996). Goal-Oriented Auditing and Logging. *ACM Transactions on Computing Systems*, accessed on 10 December 2013. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.160.9479&rep=rep1&type=pdf.

Fayyad U, Piatetsky-Shapiro G and Smyth P (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*.

Hemerly J (2013). Public Policy Consideration for Data-Driven Innovation. *Journal of Computer* **46**(6) 25-31.

Maimon O and Rokach L (2005). Introduction to knowledge discovery in databases. In: *The Data Mining and Knowledge Discovery Handbook*, edited by Maimon O and Rokach L (Springer-Verlag) New York 1–13.

Mertz CJ and Murphy PM (1996). UCI Repository of machine learning databases. University of California (Electronic version). Available: http://www.ics.uci.edu/~mlearn/MLRepository.htm (Accessed 10 December 2013).

Michael K and Miller KW (2013). Big Data: New Opportunities and New Challenges. *Journal of Computer* **46**(6) 22-24.

Patil PU (2012). Preprocessing of Web Server Log File for Web Mining. *World Journal of Science and Technology* **2**(3) 14-18, ISSN: 2231-2587.



Research Article

Pitt J, Bourazeri A, Nowak A, Roszczynska-Kurasinska M, Rychwalska A, Santiago IR, Sanchez ML, Florea M and Sanduleac M (2013). Transforming Big Data into Collective Awareness. *Journal of Computer* **46**(6) 40-45.

Ratnesh Kumar Jain, Kasana RS and Suresh Jain (2009). Efficient Web Log Mining using Doubly Linked Tree. *International Journal of Computer Science and Information Security* **3**.

Saxena M, Singh NK, Thakur SS and Kumar P (2012). A Review of Computer forensic & Logging System. *International Journal of Advanced Research in Computer Science and Software Engineering*, ISSN: 2277 128X **2**(1).

Stout and Kent (2002). Central Logging with a Twist of COTS in a Solaris Environment. *SANS Institute*. Last accessed on 10 December 2013, Available: http://www.sans.org/rr/papers/52/540.pdf.

Tallon PP (2013). Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. *Journal of Computer* 46(6) 32-38.

Valdman J (2011). Log File Analysis. *DCSE*/TR-2001-04. University of West Bohemia in Pilsen. Czech Republic.

Wigan MR and Clarke R (2013). Big Data's Big Unintended Consequences. *Journal of Computer* **46**(6) 46-53.