International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Internet Usage Analysis Using Karl Pearson's Coefficient of Correlation -A Computer Forensic Investigation

Suneeta Satpathy[1], Sateesh K. Pradhan[2], Subhasish Mohapatra[3]

[1,3]Department of Computer Science & Application, College of Engineering Bhubaneswar, BPUT, Bhubaneswar, India

[2]Post Graduate Department of Computer Application Utkal University, Bhubaneswar, India

Abstract: ***Internet has caused the distribution of illegal materials globally making criminal activity tracing difficult in the digital investigation process. In comparison to other types of forensic investigation, computer forensic examination comprises a large volume of digital data. Along with that the rapid evolution of digital technology and complexity of the devices being involved has made the criminal investigation further harder. In this paper we have used Karl Pearson's coefficient of correlation to extend our previous work to study the relationship between the estimated and predicted variables to justify the misuse of Internet for illegal activities. The prediction using correlation analysis can further simplify the digital investigation process by eliminating the conflicting information and gathering the digital evidence to be used as an expert testimony in the court of law.***

Keywords: Digital Crime, Digital Investigation, Digital Evidence, Computer Forensics, correlation, Karl Pearson's coefficient.

1. Introduction

The Internet connectivity through the ever-growing digital-infrastructure has facilitated rapid availability of information resources and databases. They have become indispensable to the smooth operation of businesses, government, and even our personal lives. Such technology has also opened the door of opportunities for the criminals to commit digital crimes known as cyber crimes [5][6]. Nonetheless, its evolution also provides equally many new sources of potential evidence of digital crimes. It is due to this increase in crimes and incidents relating to the Internet [6] and computing devices that the field of computer forensics [4] has rapidly emerged and research is being conducted into ways of improving the quality and efficiency of computer forensic investigations. Combating the complexity of forensic investigation process requires the application of tools and techniques against digital crimes. The present paper extends our previous work of forensic classification [10][11][12] to further enhance the prediction of internet usage using Karl Pearson coefficient of correlation [8]. Since correlation analysis uses simple mathematical calculations, can be easily adopted by the forensic investigators to extract patterns and anomalies from the large sets of data to add domain context information such as witness testimony, crime details into digital evidence and facilitate justification of efficient processing of evidential data.

2. Challenges in Computer Forensic Investigation

Computer Forensics uses science and technology to examine digital evidence that develops and tests theories, which can be entered into a court of law, to answer questions about events that occur [4] [5]. The purpose of such type of investigation is to find evidence related to the events under investigation and present them to the fact finder.

The major goals of computer forensics are to [4]:

- Provide a conclusive description of all digital crime activities for the purpose of complete post-attack enterprise and critical infrastructure information restoration
- Correlate, interpret, and predict adversarial actions and their impact
- Make digital data suitable and persuasive for introduction into a criminal investigative process and
- Provide sufficient legal digital evidence to allow the criminal perpetrator to be successfully prosecuted.

The evidence in computer forensic investigation termed as digital evidence [3] [4] [5] is fragile in nature and can be easily altered or destroyed. It is unique when compared to other forms of documentary evidence. As with any investigation [3] [5] to find the truth one must identify data that:

- Verifies existing data and theories (Inculpatory Evidence)
- Contradicts existing data and theories (Exculpatory Evidence)

To find both evidence types, all acquired data must be analyzed and identified. Analyzing every bit of data is a daunting task when confronted with the increasing size of storage systems [2]. Current Computer forensic investigation tools are not able to locate vital evidence within the massive volumes of data as the data generated by such tools can be misleading due to the dimensionality, complexity and amount of the data. These tools mainly focus on possible digital evidence recovery, so they are not ideal for

- Reduction of duplicate data to lessen the data for analysis
- Identifying correlations among data
- Finding and visually presenting groups of facts

Licensed Under Creative Commons Attribution CC BY

- Discovering patterns and data that may lead to reasonable predictions.

3. Problem Statement and Related Work

The Internet has become an instrumental need of all humans for gathering information and gaining knowledge. The usage of the Internet by the employees that may improve individual and consequently organizational efficiency has forced every private as well as government organizations to avail this technology as an important source of competitive strength. But deliberate misuse of the Internet for accessing and distributing illegal materials in cyber space like pirate software, child pornography materials, stolen properties, etc has become a serious problem in many organizations. Our previous work has developed a Fusion based Investigation Tool using JDL data fusion [9] by mapping the activities at different levels of Data Fusion into the steps of digital investigation process to deal with such types of crime. The investigation model also with the help of decision mining rules and statistical validation techniques [10][11] has also analyzed the interpreted information from seized hard drives. The investigation model has also enabled visualization in digital investigation process with the application of ID3 algorithm [12] for forensic classification.

Computer forensics is a synthesis of science and law [4]. At one end is the pure science of ones and zeros in which, the laws of physics and mathematics rule and at the other end, is the court-room. To get something admitted into court requires two things. First, the information must be factual. Secondly, it must be introduced with well proofed theory and methodology by a witness who can explain the facts and answer questions. While the first may be pure science, the latter requires training, experience, and an ability to communicate the science with a well defined theory [5]. Keeping in mind that the computer forensic investigation is inherently mathematical and comprises more data for an investigation than others the present work aims to extend our previous work of behavioral study of the employees in the workplaces and explore their internet usage and predict the same using Karl Pearson coefficient of correlation analysis.

4. Statistics of Correlation and Karl Pearson's coefficient of correlation

In a real scenario, two quantities are correlated if movement in one is accompanied by movements in the other. The computation concerning the degree of closeness is done by performing the correlation analysis. In other words it is an analysis of the co-variation between two or more variables. The effect of correlation is to reduce the range of uncertainty and the prediction based on correlation analysis is likely to be more valuable and near to reality. Such analysis doesn't necessarily imply causation or functional relationship though existence of causation always implies correlation [8].

There are various methods of ascertaining whether 2 variables are correlated or not. The Karl Pearson's coefficient of correlation is one of the most widely used statistical measures. It is popularly known as Pearson's coefficient of correlation, and the quantity $S_{xy} \div \sqrt{(S_{xx}S_{yy})}$ is called the correlation coefficient for X and Y, and is denoted by the symbol r [8].

$$r = S_{xy} \div \sqrt{(S_{xx}S_{yy})} \text{ ---------- eq(1)}$$

In equation 1 from the given values of X the mean of X ($\bar{x}$) can be calculated, and from this the variance of X is calculated as $(X - \bar{x})^2$, the sum of all such calculations for each individual value of X is denoted by the symbol S_{xx}. Similarly from the given values of Y the mean of Y ($\bar{y}$) can be calculated, and from this the variance of Y, which is denoted by the symbol S_{yy}. S_{xy}, is $\sum(X_i - \bar{x})(Y_i - \bar{y})$.

The full interpretation of r depends upon circumstances, when estimating the value of one variable from the value of another, the higher the value of 'r' the better the estimates. The correlation coefficient is often used as a measure of the strength of association between two variables. A value of r close to +1 or –1 indicates a strong linear association. A value close to 0 indicates a weak association. For values in between 0 and 1 the Table 1 below shows the strength of relationship between two variables [8].

Table 1: Correlation co efficient values

r value	Strength of relationship between two variables
r = 0.10 to 0.29 or r = -0.10 to – 0.29	Small relationship
r = 0.30 to 0.49 or r = -0.30 to – 0.49	Medium relationship
r = 0.50 to 1.0 or r = -0.50 to -1.0	Strong relationship

5. Data Set and Result Analysis

In our study the model of prediction and finding a relationship between two variables uses data usage of 6 people who were intentionally misusing the Internet for illegal purposes. The data collection process consists of retrieving all the files from the seized hard drives with the help of forensic toolkit FTK [1] and filtering out the picture files (.jpg, .gif, .bmp, .tiff) video files and MP3, MP4 files from it to prepare the data sets for analysis using 7 decision mining rules described in our previous work. This study further tries to establish the relationship between estimated files (X) and predicted files (Y) and prove that the classification of files designated as Y from X based on decision mining rules is valid and true with the help of correlation analysis. FTK toolkit [1] is used to retrieve picture files and mp3, mp4 and video files for all sample hard drives. These image files from each hard drive collected are designated as X. All the image files including audio and video files classified as illegal files from X based on 7 decision mining rules [10] (shown in table 2) are designated as Y. Table 3 shows the dataset for analysis.

Licensed Under Creative Commons Attribution CC BY

Table 2: Decision Mining Rules [10]

Rule no	*Rules*
1	If it is an image file created/modified/accessed early in the week (mon, tue) during 12am to 6am and 7pm to 12 am(early morning, late night) then it is suspicious.
2	If it is an image file created/modified/accessed early in the week (mon, tue) during 6am to 7pm(working hr) then it is not suspicious.
3	If it is an image file created/modified/accessed middle in the week (wed, thurs) during 12am to 6am and 7pm to 6am(early morning, late night) then it is suspicious.
4	If it is an image file created/modified/accessed middle in the week (wed, thurs) during 6am to 7pm(working hr) then it is not suspicious.
5	If it is an image file created/modified/accessed late in the week (fri, sat, sun)during 12am to 6am and 7pm to 12 am(early morning, late night) then it is suspicious.
6	If it is an image file created/modified/accessed late in the week (fri, sat, sun)during 6am to 7pm (day time working hour) then also it is suspicious.
7	But if the logical file size is large and if it is downloaded during working hours on any day of the week need investigation. Same rule is applicable for MP3 files downloaded at any time on day of the week.

Table 3: Data set for analysis

User	Total no of image files x	No of picture,videos&mp3&mp4 files Y
1	925	325
2	83	280
3	249	191
4	513	243
5	898	525
6	293	225

$\overline{X}$ is calculated as 1378/6 = 493.5 ---------- eq (2)

$\overline{Y}$ is calculated as 959/6 = 298.17 ---------- eq (3)

Table 4 shows the calculation of values of variance of X and variance of Y

Table 4: Variance calculation table

X_i	Y_i	$\overline{X}$	$\overline{Y}$	$X_i - \overline{X}$	$Y_i - \overline{Y}$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})^2$	$(X_i - \overline{X})(Y_i - \overline{Y})$
925	325	493.5	298.17	431.5	26.8	186192.25	718.24	11564.2
83	280	493.5	298.17	-410.5	-18.2	168510.25	331.24	7471.1
249	191	493.5	298.17	-244.5	-107.2	59780.25	11491.84	26210.4
513	243	493.5	298.17	19.5	-55.2	380.25	3047.04	-1076.4
898	525	493.5	298.17	404.5	226.8	163620.25	51438.24	91740.6
293	225	493.5	298.17	-200.5	-73.2	40200.25	5358.24	14676.6

$S_{xx} = \sum (X_i - \overline{X})^2 = 618683.5$ ---------- eq (4)

$S_{yy} = \sum (Y_i - \overline{Y})^2 = 72384.84$ ---------- eq (5)

$S_{xy} = \sum (X_i - \overline{X})(Y_i - \overline{Y}) = 150586.5$ ---------- eq (6)

So correlation coefficient r is calculated using the eq-1, eq-4,eq-5,eq-6 as 0.71159

6. Coefficient of Correlation and Probable Error

The probable error of the correlation coefficient helps in interpreting its value. With the help of probable error it is possible to determine the reliability of the value of the coefficient. The probable error of the coefficient of correlation is obtained as [8]

$$P.E._r = 0.6745 * (1 - r^2) / \sqrt{N} \quad \text{----- eq (7)}$$

Where r is the coefficient of correlation and N the number of pairs of observations.

Following conclusions are drawn from probability of error calculation [8].

1. If the value of r is less than the probable error there is no evidence of correlation, i.e. the value of r is not at all significant.
2. If the value of r is more than approximately six times the probable error, the coefficient of correlation is practically certain, i.e the value of r is significant.
3. By adding and subtracting the value of probable error from the coefficient of correlation we get respectively the upper and lower limit within which coefficient of correlation in the population can be expected to lie. Symbolically it is written as

$\rho = r \pm P.E.$

where ρ denotes correlation in the population.

As in our analysis r is calculated as 0.71 and total observations in the study is 6 Probable error is calculated as

$P.E._r = 0.6745 * (1 - (0.7^2)) / \sqrt{6} = 0.14$ ---------- eq(8)

The limits of the correlation in the population r ±P.E. is calculated to lie in the range 0.56 to 0.84.

According to research study instances are common where in a correlation coefficient of 0.5 or even 0.4 is obviously considered to be a fairly high degree of correlation [8] and a correlation coefficient of 0.5 means that only 25% of the variation is explained, in our study we get a correlation coefficient of 0.7 means a positive relationship exists

Licensed Under Creative Commons Attribution CC BY

between X and Y and it justifies 49% of variation is explained and there by justifies the prediction of Y value from X value using seven decision mining rules. It can also be concluded that if the no of user and Internet usage increases then quantity of misuse of such digital technology may increase, provided certain constraints remain unchanged like economic factor of the country and psychological factor of the users.

7. Conclusion

With the rapid proliferation of Internet technologies and applications, Cyber crime has become a major concern for the law enforcement community. The extensive growth of Internet and the lack of awareness and truly secure systems make it an important field of research in computer forensic. As computer forensic investigation comprises of large volume of data for investigation and can vary drastically in their level of complexity [2], each investigative process must follow a rigorous path for standardizing terminology [7], defining requirements, and supporting the development of new techniques for investigators. So in the face of information sea in the current Internet Technology, correlation analysis in digital crime investigation can help to pick out as well as predict the right crime information that reflects the crime, security requirements timely, comprehensively and accurately.

References

[1] Access Data Corporation (2005) (http://www.accessdata.com).
[2] Beebe, N., Clark, J.(2005), Dealing with terabyte data sets in digital investigations. Advances in Digital Forensics, pp. 3-16, Springer.
[3] D. Brezinski , T. Killalea (2002), Guidelines for Evidence Collection and Archiving.
[4] E. Casey (ed.)(2001) Handbook of Computer Crime Investigation, Academic Press.
[5] E. Casey (2004), Digital Evidence and Computer Crime, 2nd Edition, Elsevier Academic Press.
[6] H Lipson (November 2002), Tracking and Tracing Cyber Attacks: Technical Challenges and Global Policy Issues (CMU/SEI-2002-SR-009), CERT Coordination Center.
[7] Meyers, M. ,Rogers, M. (2004), Computer forensics: the need for standardization and Certification, International Journal of Digital Evidence, vol. 3, no. 2.
[8] Statistical Methods by Dr. S.P Gupta, Sultan Chand & Sons Educational Publishers.
[9] Satpathy Suneeta, Sateesh Pradhan, and B.N.B Ray (2010), A Digital Investigation Tool based on Data Fusion in Management of Cyber Security Systems, International Journal of IT & Knowledge Management, volume-3, Number 2 pp 561- 565, (http://www.csjournals.com/IJITKM/PDF%203-1/77.pdf).
[10] Satpathy Suneeta, Sateesh Pradhan, and B.N.B Ray (2011), Rule based Decision Mining with JDL Data Fusion model for Computer Forensics: A Hypothetical Case Study , International Journal of Computer Science and Information Security, volume-9, Number 12 pp 93-100, http://sites.google.com/site/ijcsis .
[11] Satpathy Suneeta, Sateesh Pradhan, and B.N.B Ray (2012), Application of data fusion methodology for computer forensics dataset analysis to resolve data quality issues in predictive digital evidence, International Journal of Forensic Computer Science, volume-7, Number 1 pp 16- 23, http://dx.doi.org/10.5769/J201201002.
[12] Satpathy Suneeta, Sateesh Pradhan, and B.N.B Ray (2014), A Decision Driven Computer Forensic Classification Using ID3 Algorithm, Intelligent Computing, Communication and Devices, Volume 309, 2015, pp 367-376, DOI 10.1007/978-81-322-2009-1_42.

Licensed Under Creative Commons Attribution CC BY