Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

An investigation of Zipf's Law for fraud detection (DSS#06-10-1826R(2))

Shi-Ming Huang [a], David C. Yen [b,*], Luen-Wei Yang [a], Jing-Shiuan Hua [c]

[a] Department of Accounting & Information Technology, National Chung-Cheng University, Chia-Yi, Taiwan, ROC
[b] Department of DSC & MIS, Miami University, Oxford, OH 45056, United States
[c] Department of Information Management, National Chung-Cheng University, Chia-Yi, Taiwan, ROC

ARTICLE INFO

Article history:
Received 5 October 2006
Received in revised form 4 May 2008
Accepted 25 May 2008
Available online 16 July 2008

Keywords:
Fraud detection
Zipf's Law
Benford's Law
Quasi-experiment research
Misclassification cost matrix

ABSTRACT

Fraud risk is higher than ever before. Unfortunately, many auditors lack the expertise to deal with the related risks. The objectives of this research are to develop an innovative fraud detection mechanism on the basis of Zipf's Law. The purpose of this technique is to assist auditors in reviewing the overwhelming volumes of datasets and identifying any potential fraud records. The authors conducted Quasi-experiment research on the KDDCUP'99 benchmark intrusion detection dataset to verify the performance of the proposed mechanism. The simulation experimental results demonstrate that Zipf Analysis can assist auditors to locate the source of suspicion and further enhance the resulting audit processes.

1. Introduction

Fraud risk is higher than ever before. According to the results of KPMG's Fraud Survey of 2003, organizations are reporting more experiences of fraud than in prior years [16]. In 2003, 75% of surveyed companies reported that they experienced an instance of fraud, an increase of 13% as compared with 1998. Furthermore, Ernst & Young's Global Survey pointed out that the main contributing factors to the prevalence of fraud are the growing complexity of organizations and systems, changes in business processes and activities, enormous and ever-expanding volumes of transaction data, outdated and ineffective internal controls and so on [11]. Complex organizations and transactions lead to increased opportunity for subjective interpretation, borderline disclosure and even misrepresentation.

Auditors must take fraud prevention and awareness seriously. The main reason is that the occupational fraud issue against organizations is a costly business problem. The economic impact can be significant: up to 6% of organizations' revenues may be lost annually as a result of fraud and abuse [3].[1] Within the United States, this translates into losses of approximately $660 billion. Furthermore, the true cost of fraud goes beyond the financial loss to the impact on reputation, diversion of management focus, and loss of morale and trust within teams [11]. However, recent corporate scandals clearly indicate the potential for fraud abuse and many auditors lack the skills and expertise to deal with the related risks.

It is clear that there is a critical need to implement some technology solutions to auditing areas, especially the data analysis technique [17]. Proper data analysis is critical as a means of allowing auditors to streamline audit processes, bring fraudulent activities to light before they result in critical losses, minimize financial losses, and ensure compliance with business rules and external regulatory requirements, such as SAS 410 and SOX.

In this research, the authors propose an innovative mechanism of analytical review procedure, named “Zipf Analysis” that utilizes the conception of Zipf's Law to facilitate the systematic construction of a fraud detection model. For the purpose of building such fraud detection systems, we examine the properties of Zipf's Law and evaluate its capability of discriminating abnormal records from a simulation experiment

* Corresponding author. Tel.: +1 513 529 4827; fax: +1 513 529 9689.
E-mail addresses: smhuang@mis.ccu.edu.tw (S.-M. Huang), yendc@muohio.edu (D.C. Yen), ccuait404@gmail.com (L.-W. Yang), jshua@mis.ccu.edu.tw (J.-S. Hua).

[1] ACFE, Report to the nation, http://www.acfe.com/fraud/report.asp.

0167-9236/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2008.05.003

on the KDDCUP'99 intrusion detection dataset, and revising the fraud detection system to provide a higher level of accuracy and efficiency.

This paper is organized as follows: Section 2 reviews related literatures about analytical procedures: Benford's Law and Zipf's Law. Section 3 describes the systematic mechanism of the fraud detection model on the basis of Zipf's Law. Sections 4 and 5, present a simulation experiment for the system implementation and evaluation of the proposed mechanism by comparing other fraud detection methods. Finally, the conclusions of this study are provided in Section 6.

2. The problem and related work

Fraud by nature is composed of the following three categories: intentional illegal act, the concealment of that act, and deriving a benefit from that act [5,8]. In recent years, deceptive behaviors have been transformed and have emerged into new paradigms, such as money laundering, etc. Fortunately, it is indeed the case that some enterprise-wide fraud can be systematically detected with the assistance of emerging information technologies. Conventionally, data analysis approaches using power laws have been adopted as a feasible artifact to offer valuable assistance to fraud detection in computer auditing operations.

Benford's Law is actually one of the most commonly utilized power laws for fraud detection. The underlying concept Benford's Law states that in the lists of numbers obtained from real-life data resources, the distribution of leading digit is a long-tail distribution [4]. To this end, many prior researchers have proved that auditors could improve their performance by conducting analytical procedures before utilizing substantive tests [10,12,25,32]. For example, the study of Reed and Pence [25] examines the appropriateness of Digital Analysis as an analytical review procedure to determine the possibility of financial statement fraud. They indicate that using Digital Analysis in analytical procedures can be more effective or efficient than tests of details in reducing detection risk for specific financial statement assertions. In addition, the Statement of Auditing Standards No. 410 also suggested that auditors should apply analytical procedures at the planning stage to assist in understanding the business and in identifying areas of potential risk. On the other hand, some studies in the accounting area have also suggested that Digital Analysis could be used as an analytical review procedure to assist auditors in identifying any possible fraud issues. However, there are some limitations while using Benford's Law for fraud detections [30]. The restrictions noted by the author include the following:

1) Some populations of accounting-related data do not conform to a Benford distribution [10] [18,19]. For instance, we found that in some businesses, numbers (especially monetary amounts) in transactions are not a natural set. For example, when evaluating a data file of travel claims you might find that the first-two-digit combination of 24 appears more often than expected with Benford's Law. This might happen if the company has a policy that stipulates that reimbursement claims for $25 and above must be supported with receipts — so travelers claim a lesser amount, such as $24.95. Therefore, it is hard to use Benford's Law when the amounts are mostly identical.
2) Using Digital Analysis may generate too many cases to review [30]. As an example, in a database containing 250,000 records, one expects that approximately 12,250 records will start with 10. Should one find 15,000 numbers starting with 10, there is a doubt, and we still have to check 15,000 transactions. Therefore, we understand that some other filters are necessary to narrow down these sets of data to a manageable size and to reduce some noisy data.
3) When execute analytical procedures with Benford's Law, auditors only can use the data of digital shape for analysis. However, Statements on Auditing Standards, SAS 410, suggests that analytical procedures should be designed to consider each kind of different data shape simultaneously and identify the interrelation of financial and non-financial data.

Although Digital Analysis using Benford's Law had been proven to facilitate auditor review on overwhelming volumes of data and transactions, there are many drawbacks from using this particular method. After completing related literature review we found another power law by the name of Zipf's Law. The basic concept of Zipf's Law is that the frequency of the word occurrence in an article in fact furnishes a useful measurement and hence, management of word significance: The product of frequency of the use of words, f, and the rank order, r, is approximately constant [39]. Many scholars believe that Benford's Law is a special case of Zipf's Law [21]. According to Zipf's Law, we believe we can use Zipf's Law to verify the frequency of string or date fields in records. Therefore, this research aims to construct an analytical procedure mechanism on the basis of Zipf's Law and further test its performance of detecting anomalies by comparing with other fraud detection algorithm. The following Table 1 summarized the comparison between Digital Analysis using Benford's Law and Zipf's Law, accordingly.

The concept of Zipf's Law has also been adopted in the area of Information Retrieval. Information Retrieval (IR) typically involves problems inherent to the collection process for a corpus of documents, and then provides functionalities for users to find a particular subset of it by constructing queries [13]. Basically, the idea of IR implementation revolves around an attempt to systematically extract information from available document corpora, and then utilize it to determine whether or not each document is relevant to a particular request [28]. Among the various IR models, the Vector Space Model (VSM) is one of the most popular and successful approaches for modeling in the IR associated environment

Table 1
Comparison between Benford's Law and Zipf's Law

Benford's Law	Zipf's Law
They are both derived from Nature Laws.	
They can be used to handle disaggregated account level data.	
They both follow the principle of Power Law.	
Shows relationship between digit and frequency.	Shows relationship between rank and frequency.
Numeric attributes are required	No pre-requirements defined for type of attributes.
Applied for fraud detection already	Under review as a potential tool for fraud detection