Information Systems 44 (2014) 22–33

Contents lists available at ScienceDirect

Information Systems

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

Outlier detection in audit logs for application systems

H.D. Kuna [a],*, R. García-Martinez [b], F.R. Villatoro [c]

[a] Computer Science Department. School of Sciences, National University of Misiones, Félix de Azara 1552, Zip Code: N3300LQH, Posadas, Misiones, Argentina
[b] Information Systems Research Group. Productive and Technological Development Department, National University of Lanús, Argentina
[c] Language and Computer Science Department, University of Malaga. Spain

ARTICLE INFO

Article history:
Received 15 January 2014
Received in revised form
5 March 2014
Accepted 8 March 2014
Available online 26 March 2014

Keywords:
Data mining
Systems audit
Outlier detection

ABSTRACT

An outlier is defined as an observation that is significantly different from the other data in its set. An auditor will employ many techniques, processes and tools to identify these entries, and data mining is one such medium through which the auditor can analyze information. The enormous amount of information contained within transactional processing systems' logs means that auditors must employ automated systems for anomalous data detection. Several data mining algorithms have been tested, especially those that deal specifically with classification and outlier detection. A group of these previously described algorithms was selected for use in designing and developing a process to assist the auditor in anomalous data detection within audit logs. We have been successful in creating and ratifying an outlier detection process that works in the alphanumeric fields of the audit logs from an information system, thus constituting a useful tool for system auditors performing data analysis tasks.

Contents

* Corresponding author.
E-mail address: hdkuna@unam.edu.ar (H.D. Kuna).

http://dx.doi.org/10.1016/j.is.2014.03.001
0306-4379/© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Information has become a key resource for all organizations, and Information Technologies (IT) have become increasingly widespread and are now firmly rooted in every aspect of organization management. Organizations have made gathering quality information to aid both in everyday activities and in decision making an important goal. Corporations must have formal processes in place to guarantee the legality, security and quality of their information.

Systems auditing is composed of a series of tasks aimed at ensuring that all information systems within an organization function properly and at providing the basis that enables corporations to fulfill their strategic objectives. The list of good practices developed by the ISACA (Information Systems Audit and Control Association) [8] within its COBIT framework provides guidelines to aid organizations in achieving their corporate IT governance and management objectives.

Audit logs contain records of every operation carried out within a software information system and play a key role in guaranteeing that each organization's procedures and regulations are observed. Finding anomalies through manual queries or analyses of the audit logs' stored data requires highly trained staff and a significant expenditure of man-hours. An outlier is an observation [16] suspected of originating from an alternative input mechanism due to its distinguishing features. Detecting these outliers in audit logs is extremely useful, as their existence can provide the auditor with crucial information, but manual searches would be too time-consuming due to the huge amounts of data found in these logs.

Automated mechanisms, and data mining in particular, are of great use in this field because of their ability to detect patterns and non-obvious correlations among different pieces of data.

Data mining, described as the process of intelligently extracting useful, non-apparent information from databases, has been utilized widely in systems auditing [7].

Some data mining techniques focus on outlier detection. Anomalous data may stem from the software systems' operating noise, and detecting these entries should be of paramount importance for the system auditor. Anomalous data detection in transactional software audit data logs is particularly important, as the risks posed by these anomalies may threaten the system's proper operation.

Real databases contain anomalies related to different causes, including errors in data collection, errors in the information systems, probable malicious actions, and so on. In the particular case of the audit logs of application systems, anomalies can occur due to operations carried out within the system that are not considered common, errors in the information system recording the audit logs, modifications to the audit log, and so on. In all cases, these audit trails signifying the detection of anomalous values must be eventually analyzed by the auditor because the underlying cause of these outliers may imply a risk to the security or quality of the data.

This paper aims to introduce a process that employs data mining techniques to automate outlier detection in system audit logs that include alphanumeric data. Automated detection can allow an auditor to detect hints of anomalous activities, which will most likely require closer scrutiny. This system must also be usable by system auditors, even if they are not experts in data mining.

1.1. Related work. data mining in systems auditing

Computer-Assisted Auditing Techniques (CAATs) make it possible to use computers as part of the auditing process. Data mining is one of the available techniques, but there are other options, namely, the following:

- Data analysis software.
- Network security assessment software.
- Assessment software for operating systems and database management systems.
- Software and source code testing tools.

Many works have been published on intrusion detection in the log files of network operating systems, although fewer studies exist on management systems logs. [22,31,38] Outlier detection applications may also be found within databases [40,41].

The extant papers define a taxonomy for anomalies found through outlier detection [7], while some other papers make mention of work conducted on fraud detection for credit cards [4,37] and cellular phones [12].

Clustering is a data mining technique that may be employed for outlier detection. This strategy consists of unsupervised learning, during which data are automatically assigned to one of several clusters according to certain shared characteristics. A tuple is more likely to be tagged as an outlier the further it falls from the rest of the sample. Several clustering techniques are available, including the following:

- Hierarchical clustering, which produces a hierarchy of clusters within the dataset. This technique's results are usually presented in a dendrogram.
- Partitioning methods, in which the dataset is successively partitioned. Objects are clustered into different groups and therefore each object's deviation from the cluster's centers must be kept to a minimum.