Data Mining in Computer Auditing

Teh Ying Wah , Mustaffa Kamal Mohd Nor , Zaitun Abu Bakar and Lee Sai Peck
University of Malaya, Kuala Lumpur, Malaysia

tehyw@fsktm.um.edu.my mustaffa@fsktm.um.edu.my zaitun@fsktm.um.edu.my saipeck@fsktm.um.edu.my

Abstract

In this paper, we first introduce the readers about the main function of a computer auditor. This is followed by a description of auditing the usage of stationeries in the Faculty of Computer Science and Information Technology, University of Malaya. It is a very time consuming process to audit all stationeries. Therefore, we introduce the data mining techniques to help us find the relevant stationeries. We use this information to recommend purchasers to purchase relevant items together in order to achieve efficiently in purchasing stationeries process.

Keywords: Web-Based Environment, Personalisation, Data Mining, Computer Auditing and Association Rules.

Introduction

Traditionally, the main function of a computer auditor was the operational review of application systems. However, more and more audit departments are adopting a more integrated approach where the computer audit specialists will work with general internal auditors to review business processes of which only one part is supporting the application systems. These reviews can be done as the system is developed, as a post-implementation review, or as a regular audit of the application. Whichever stage of audit review is being carried out, the auditor is looking for assurance that the application provides an adequate degree of control over the data being processed. The level of control expected for a particular application is dependent on the degree of risk involved in the incorrect or unauthorised use for processing the data. The sensitivity of that data must therefore be determined at an early stage [1].

It is very difficult to be specific about the actual control mechanisms that the computer auditor should look for. These will vary depending on the underlying technology used to power the systems. However, some generic principles are offered for guidance. The computer auditor needs to make very subjective judgments when reviewing an actual system to determine if the particular mechanisms implemented are effective [1].

Computer Auditing

Computer audits proactively identify and evaluate information technology related control weaknesses and risks, by focusing on the availability, confidentiality and integrity of Computer Information Systems (CIS). The controls in a CIS environment include manual procedures as well as procedures in computer programs. The control environment consists of management and independent controls [2].

The following definitions of the key controls are given for the sake of clarity [2]:

(a) **Management and independent controls** are direct controls which are performed by persons independent of the processing thereof, to detect errors or irregularities which may have occurred before or during processing and may not have been prevented by processing controls. They typically comprise manual reviews, analysis, comparisons and reconciliations.

(b) **General controls** support the effectiveness of management and independent controls as well as processing controls. Their operation is often essential for the effectiveness of application controls and they typically comprise:

- Controls over planning, policies, procedures and standards pertaining to the Computer Information Systems (CIS) environment.
- Controls over physical, logical and network security.
- Controls over program change controls, for both maintenance of existing systems and development of new systems.
- Controls over operational procedures, backups and business continuity.
- Controls over database management.

Controls over the CIS organization.

(c) **Application controls** can be manual or computerised and can be applied either to individual transactions or similar batches designed to prevent or detect errors or irregularities occurring in the early stages of processing or immediately thereafter. These include controls over processing and computer data designed to provide reasonable assurance that:

(i) data is entered correctly;

(ii) transactions are properly processed by the computer;

(iii) transactions are not lost, added, duplicated or improperly changed;

(iv) processing errors are identified and corrected on a timely basis; and

(v) transactions are not reported incorrectly.

The overall purpose of these controls is to reduce expected losses from unlawful events that can occur in a system.

The auditor's duty is to determine whether controls are in position and working effectively to prevent the illegal events that might happen within a system. The malfunctioning of these controls will result in unacceptable losses and unreliable system that provide the auditor with unnecessary data and information.

To explain the process of auditing, we start with auditing the usage of stationeries in the Faculty of Computer Science and Information Technology, University of Malaya.

There are many types of stationery in the faculty. What are the relevant stationeries for auditing? This issue motivates us to discuss further. It is a very time consuming process to audit all stationeries. Therefore, we introduce the data mining techniques to help us find the relevant stationeries.

Data Mining

One of the popular citations defines data mining as the non-trivial extraction of hidden, previously unidentified, and potentially valuable knowledge from data [7].

Another definition is that data mining is a variety of techniques such as neural networks, decision trees or standard statistical techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in areas such as decision support, prediction, forecasting, and estimation [8].

Hand [6] states that data mining is *a new discipline lying at the interface of statistics, data base technology, pattern recognition, and machine learning, and concerned with secondary analysis of large databases in order to find previously unsuspected relationships, which are of interest of value to their owners*.

To researchers and practitioners, data mining and knowledge discovery in databases are terms often used interchangeably. Data mining is the process of finding trends and patterns in data. The objective of this process is to sort through large quantities of data and discover new information [3]. Dr. Penzias, a Nobel Prize winner interviewed in ComputerWorld [3] in January 1999, comments: *Data mining will become much more important, and companies will throw away nothing about their customers because it will be so valuable. If you're not doing this, you're out of business.*

Research Objectives

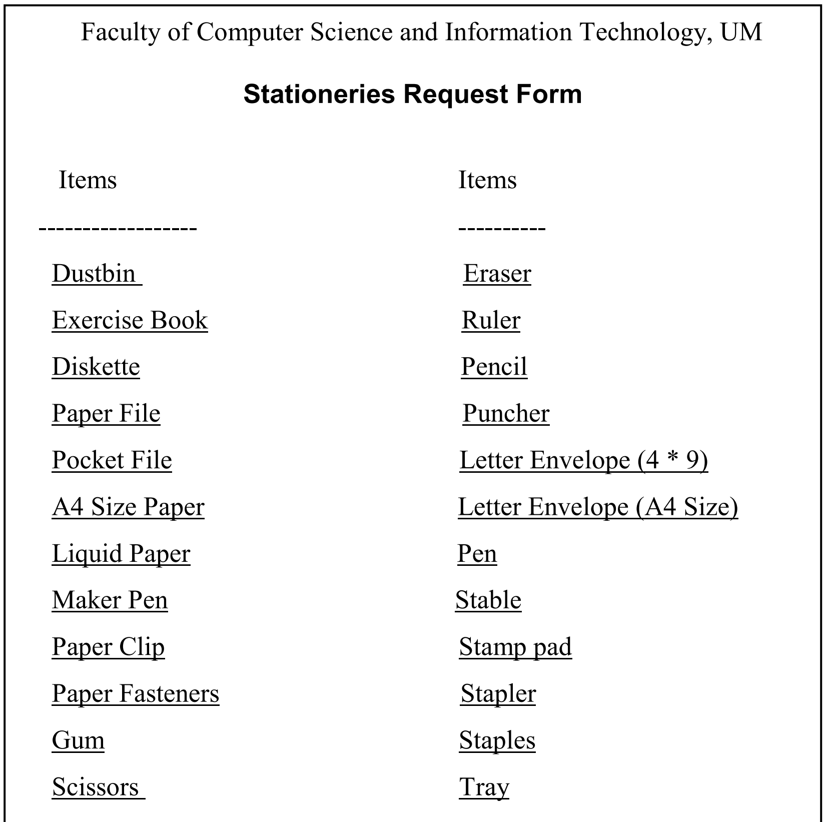
Relevant access is especially important for auditors. The reason is that an auditor is given short time frame to check huge amount of stationeries. Recognising the important role of the right selected stationeries in auditing processing, the question in the following is posed:

- What are the important criteria for auditors to improve the auditing process?

The objective of this study is to apply data mining techniques in the auditing process to generate relevant stationeries.

Association Rule Algorithms

We create a web page to let faculties or staff to request stationeries through the Intranet as shown Figure 1.



Faculty of Computer Science and Information Technology, UM

Stationeries Request Form

Items

Dustbin
Exercise Book
Diskette
Paper File
Pocket File
A4 Size Paper
Liquid Paper
Maker Pen
Paper Clip
Paper Fasteners
Gum
Scissors

Items

Eraser
Ruler
Pencil
Puncher
Letter Envelope (4 * 9)
Letter Envelope (A4 Size)
Pen
Stable
Stamp pad
Stapler
Staples
Tray

Figure 1: A stationery's web page contains 24 Items

Assume that there are only five items being requested throughout a year as shown in Table 1:

Time	Items Involved
T_1	Exercise Book, Pencil
T_2	Eraser, Pen , Exercise Book, Tray
T_3	Eraser, A4 size paper, Pen, Pencil
T_4	Eraser, Exercise Book, Pen, Pencil
T_5	A4 size paper, Exercise Book, Pen, Pencil

Table 1: Log file of Items being accessed

The idea of association rule mining algorithms begins from the study of market-basket data. The problem assumes we have some large number of items ("pen", "pencil"). Staffs fill their market baskets with some subset of items and we get to know what items staff request together. Auditors use this information to recommend purchasers to purchase relevant items together.

There are three main elements that make up the association rule mining algorithm such as associate rule, support and confidence.

Association Rule: $X \Rightarrow s,\ \alpha\ \ Y$

Support : $s = \sigma\,(X\ U\ y) / |T|$

Confidence : $\alpha = \sigma\,(X\ U\ y) / \sigma\,(X)$

Where X = item or itemset

Y = item or itemset

$|T|$ = total number of Transactions

σ = occurrence of item or itemset

s = support

α = confidence

Example:

$\{pen, pencil\} \Rightarrow s,\ \alpha$ Eraser

$s = \sigma$ (pen, pencil, eraser) / Total Number of Transactions = 2/5 = 0.4

$\alpha = \sigma$ (pen, pencil, eraser) / σ (pen, pencil) = 0.66

```
Open the access log file (as shown Table 1) as input
Open one attribute file with count as output
Open one attribute file with minimum support as output
Do
     Read a record from the access log file
    Keep count for respective attributes and store them into the   respective array of attributes
While (not end of file)
Do
   Read data from the array of attributes
   Store it into the count of Table 2
  If count of Item >= minimum support then
    Store it into one attribute with minimum support file (as shown Table 3)
  End if
While (not end of array)
```

Figure 2: Algorithm for Generating a Table for One Attribute Item, Count and Minimum Support

Assume that the minimum support is 3. The algorithm in Figure 2 demonstrates how to process the access log file and generate a table of one attribute item, count and minimum support.

Table 2 and Table 3 are the output generated by the algorithm given in **Figure 2.**

Item	**Count**
Exercise Book	4
A4 size paper	2
Pencil	4
Eraser	3
Pen	4
Tray	1

Table 2: One Attribute Item and Count

Item	**Minimum Support**
Exercise Book	4
Pencil	4
Eraser	3
Pen	4

Table 3: One Attribute Item and Minimum Support Having at least 3

The algorithm in Figure 3 demonstrates how to process Table 3 and generate a table of two-attributes itemset, count and minimum support.

```
Open Table 3 as input
Open two-attribute file with count as output
Open two-attribute file with minimum support as output
For I = 1 to number of items in Table 3
    For J = I to number of items in Table 3
        Union  item of I with  item of J + 1 and store it into Table 4
    End for J
End for I
 Do
     Read a record from Table 4
    Keep count for respective itemsets and store them into the   respective array of attributes
While (not end of file)
Do
   Read data from the array of itemsets
   Store it into the count of Table 4
  If count of Itemset >= minimum support then
    Store it into two attributes with minimum support file (as shown Table 5)
  End if
While (not end of array)
```

Figure 3: Algorithm for Generating a Table for Two Attributes Itemset, Count and Minimum Support

Table 4 and Table 5 are the output generated by the algorithm given in **Figure 3.**

Itemset	**Count**
{Exercise Book, Pencil}	3
{Exercise Book, Eraser}	2
{Exercise Book, Pen}	3
{Pencil, Eraser}	2
{Pencil, Pen}	3
{Eraser, Pen}	3

Table 4: Two Attributes Item sets and Count

Itemset	**Minimum Support**
{Exercise Book, Pencil}	3
{Exercise Book, Pen}	3
{Pencil, Pen}	3
{Eraser, Pen}	3

Table 5: Two attributes item sets and Minimum Support Having at least 3

The algorithm in Figure 4 demonstrates how to process Table 5 and generate a table of two attributes itemset, count and minimum support.

```
Open Table 5 as input
Open three attributes file with count as output
Open three attributes file with minimum support as output
 For I = 1 to number of itemsets in Table 5
    For J = I to number of itemsets in Table 5
        Union  itemset of I with  itemset of J + 1 and store it into Table 6
    End for J
End for I
Do
     Read a record from Table 5
    Keep count for respective itemsets and store them into the   respective array of attributes
While (not end of file)
Do
   Read data from the array of itemsets
   Store it into the count of Table 6
  If count of Itemset >= minimum support then
    Store it into three attributes with minimum support file (as shown Table 7)
  End if
While (not end of array)
```

Figure 4: Algorithm for Generating a Table for Three Attributes Itemset, Count and Minimum Support

Table 6 and Table 7 are the output generated by the algorithm given in Figure 4.

Itemset	Count
{Exercise Book, Pencil, Pen}	3
{Pencil, Pen, Eraser}	2

Table 6: Three attributes item sets and Count

Two attributes item sets	Minimum Support
{Exercise Book, Pencil, Pen}	3

Table 7: Three attributes item sets and Minimum Support Having at least 3

We will select {Exercise Book, Pencil, Pen} as our candidate choice.

Conclusion

We use data mining technique such as association rule in our study. We started with collecting data from the log file for a user. Based on the log file, we are able to discover the frequently access attributes by using association rule algorithms. Then, we generate three attributes item set. Based on the three attributes item set, we use this information to recommend purchasers to purchase relevant items together.

References:

1. Oliphant, A., *An Introduction to Computer Auditing*. Available from World Wide Web: http://www.itaudit.org/forum/newitauditor/f403na.htm.

2. Office of the Auditor-General *Computer auditing : Background*, 2001. Available from World Wide Web: http://www.agsa.co.za/Audicom/B6%20-%20Information_vol1_1998.html

3. Groth R., (2000). *Data Mining Building Competitive Advantage,* Prentice Hall.

4. Agrawal, R. Srikant., (1994), ``Fast Algorithms for Mining Association Rules'', Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile.

5. R. Weber, (1999). *Information Systems Control and Audit*, Prentice Hall.

6. D. Hand, (1998). Data Mining: Statistics and More? The American Statistician.

7. PAKDD, (2002). Toward the Foundation of Data Mining. 2002. Available from World Wide Web: http://www.mathcs.sjsu.edu/faculty/tylin/pakdd_workshop.html. Last modified: February 20 2002.

8. GeneCards, (2002). Knowledge Discovery In Biology and Medicine. Available from World Wide Web: http://bioinfo.weizmann.ac.il/cards/knowledge.html. Last modified: February 26 2002.

Biographies

Teh Ying Wah is a Lecturer at the Faculty of Computer Science and Information Technology in University of Malaya.He holds a B.Sc., M. Sc. in computer science from Oklahoma City University, USA. He is currently carrying out research in data mining, data warehouses and E-Commerce.

Mustaffa Kamal Mohd Nor holds a Master of Science in Computing from the University of Northumbria at Newcastle Upon Tyne, UK. in 1997. He is presently a lecturer specializing in Information Systems, Computer Auditing, Web-Based Applications and Strategic Management/ Planning in IS. He also has some experiences in computerized inventory systems and online retailing.

Zaitun Abu Bakar obtained her PhD. in Computer Science from the University of Malaya in 1999. Curently, she is an Associate Professor at the Faculty of Computer Science and Information Technology, University of Malaya. Her research areas include e-learning, e-commerce and IT in government. She has published a number of papers related to those areas.

Lee Sai Peck is currently an associate professor at Faculty of Computer Science & Information Technology, University of Malaya. She obtained her Master of Computer Science from University of Malaya in August 1990, her Diplôme d'Études Approfondies (D. E. A.) in Computer Science from University of Pierre et Marie Curie (Paris VI) in July 1991 and her Ph.D. degree in Computer Science from University of Panthéon-Sorbonne (Paris I) in July 1994. Her current research interests include Software Engineering, Object-Oriented (OO) Methodology, Software Reuse and Framework-based Development, Information Systems and Database Engineering, OO Analysis and Design for E-Commerce Applications and Auction Protocols. She has published a number of research papers in several computer science journals as well as in local and international conferences. She is a member of IEEE Computer