Minimizing Security Risk Areas revealed by Data mining

M. LOOCK[1], J.H.P. ELOFF[2]
[1]*loockm@unisa.ac.za*
Department of Computer Science and Information Systems
University of South Africa, Pretoria, South Africa
Tel: +27 12 429-6122 Fax: +27 12 429-4868

[2]*eloff@rkw.rau.ac.za*
RAU Standard Bank Academy for Information Technology
Rand Afrikaans University, Johannesburg, South Africa
Tel: +27 11 489-2842 Fax: +27 11 489-2138

Key words: Information security; data mining; knowledge discovery; data warehouse; mining database.

Abstract: The main reason for doing Data Mining is for knowledge discovery which means Data Mining finds 'knowledge' that is otherwise hidden by large volumes of data. When mining for hidden knowledge causes a potential security risk - if the knowledge is hidden, how do we know that a security risk exists? Interesting is where the security of individual data items is not a concern, but there may be patterns in the mined data that pose a security risk. This article will present an example where data mining causes an information security risk under these circumstances. The question will then be asked: What can one do about information security risk problems that surfaces while doing 'normal and legal' data mining? No answer will be given at this stage but in stead, the initial stages of this research will be explained namely looking at the Information Security Services and how do they cross reference with the Data Mining Process.

1. INTRODUCTION

The aim and propose of Data Mining is knowledge discovery which means Data Mining finds 'knowledge' that is otherwise hidden by large volumes of data. This now points to a potential security risk; if the knowledge is hidden, how do we know that a security risk exists? Interesting is where the security of individual data items is not a concern, but there may be patterns in the mined data that pose a security risk.

This article will present an example where data mining causes an information security risk under these circumstances. The question will then be asked: What can one do about information security risk problems that surfaces while doing 'normal and legal' data mining? No answer will be given at this stage but in stead, the initial stages of this research will be explained.

2. DATA WAREHOUSES AND DATA MINING

Turning the large volumes of data, data about everything around us, into information and knowledge is a necessity. The information and knowledge now gained can be used for applications such as business management, market analysis, and science exploration.

Some time ago, during the late 1980's, an architecture for databases started to grow namely data warehouses. A data warehouse refers to a database that is maintained separately from an organization's operational databases. It collects information about subjects across an entire organization, which means its scope is enterprise-wide. It allows for the integration of a mixture of application systems and it supports information processing by providing a reliable platform of consolidated historical data for analysis. A short and comprehensive definition for a data warehouse is: 'A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process' [Inmon, 1996]. For the purposes of this definition subject-oriented means organized around major subjects such as customer or sales; integrated means combining multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records; time-variant means data are stored to provide information from a historical perspective (e.g., the past 6-12 years); non-volatile means a data warehouse is a physically separate store of data deducted from the application data found in the operational environment and for this reason no transaction processing, recovery or concurrency control mechanisms are required except for two operations in data accessing namely initial loading of data and access of data.

To understand data warehouses better, the following is also important. Data warehouse technology includes, amongst others, data cleaning, data integration, and On-Line Analytical Processing (OLAP). OLAP is analysis techniques with functionalities such as summarization, consolidation, and aggregation, as well as the ability to view information from different angles. Although OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis, such as data classification, clustering, and the characterization of data changes over time. Data cleaning and data integration is done when constructing a data warehouse, which can be viewed as an important pre-processing step for data mining. This now brings us to the next question namely: What is data mining?

3. DATA MINING FUNDAMENTALS

The concept of data mining and what it entails now needs some explanation. Data mining is a logical concept built on already existing fields, techniques, and tools and on the lowest level it is applied to data. Data mining refers to extracting or 'mining' knowledge from large amounts of data. This mining process uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. Sometimes data mining is treated as a synonym for another popular term, Knowledge Discovery in Databases, or KDD [Han J & Kamber M, 2001]. Another view is that data mining is simply an essential step in the process of knowledge discovery in databases.

Knowledge discovery as a process is pictured in Figure 1 and consists of an interactive sequence of the following steps:

1. Define the business problem (Make a clear statement of your objectives, which includes a way of measuring the results of the knowledge discovery process. The business problem may also include a cost justification.).
2. Build a mining database (1st out of 3 Data Preparation steps).
    a. Data collection (identify possible multiple sources of data that will be mined)
    b. Data description (describe contents of each file or database table).
    c. Data selection (this is a gross elimination of irrelevant or unneeded data).
    d. Data cleaning and data quality assessment (to remove noise and inconsistent data).
    e. Data integration and consolidation (combining data from different sources into a single mining database).

   f. Metadata construction (this is a database about the mining database).
   g. Load the mining database.
   h. Maintain the mining database (needs to be backed up, performance monitored, and reorganized).
3. Explore the data (identify most important fields in predicting an outcome, and determine which derived values may be useful). (2nd out of three Data Preparation steps).
4. Prepare data for modelling. (3rd out of three Data Preparation steps).
   a. Select variables (delete irrelevant variables to increase time building a model).
   b. Select rows (delete irrelevant variables to increase time building a model).
   c. Construct new variables (some variables that extend over a wide range may be modified to construct a better predictor, such as using the log of income instead of income).
   d. Transform variables (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

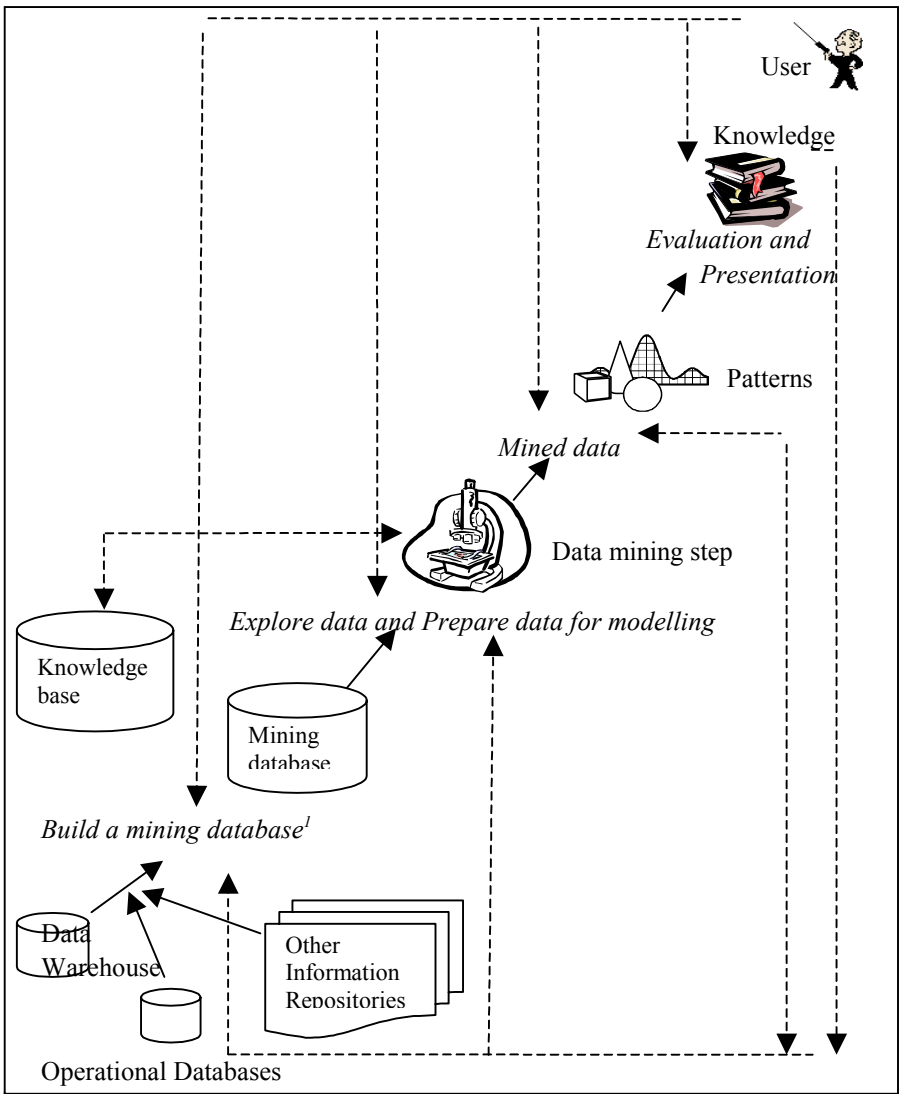

Figure 1: Data mining as a step in the process of knowledge discovery

[1] Build a mining database = data collection, data description, data selection, data cleaning, data integration, metadata construction, load mining database and maintain mining database

5. Data mining step (an essential process where intelligent methods are applied in order to extract data patterns and build models in solving business problems). This step may interact with the user or a knowledge base.
6. Pattern evaluation and interpretation (to identify the truly interesting patterns representing knowledge, based on some interestingness measures).
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

From this one can see the truth in data mining being a step in the knowledge discovery process but for the purposes of this research the term data mining will be used for the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories.

4. INFORMATION SECURITY ISSUES

It is now appropriate to ask the question: 'Where does Information Security fits into all of this?' When following the Data Mining process as described earlier, there is a noticeable collection of Information Security issues that cannot be ignored. Some of these issues will now be demonstrated by means of an example based on work done by Chris Clifton [Clifton, 1998].

Example: Risks posed by patterns in the data of top business enterprises, as mined by an external travel agent - Prediction of sensitive information.

Suppose, for the top business enterprises in South Africa, major corporate announcements required a face-to-face meeting of top and senior management from various locations throughout the country. In addition, negative announcements required participation of top and senior management as well as senior public relations staff throughout the country. Travel records (likely made available to an external travel agent) could then be used to predict the occurrence ('When will the next major corporate announcements be made?') and type ('positive or negative announcements') of corporate announcements. Here the individual travel records are not a concern, but their correlation with past announcements poses a risk.

The external travel agent will be able to answer the following questions with great ease:

'When will the next major corporate announcements be made?'

After all the top, senior management, and senior public relations staff has travelled to head office.

'Will the next corporate announcement be positive or negative?'

It will be positive if only the top and senior management was brought to head office but it will be negative if the senior public relations staff was also brought to head office.

This emphasizes the statement that one should look at the Data mining process and establish the stages that poses an information exposure risk. To be able to establish this risk we must first look at Information Security, as defined in the ISO 7498-2 standard, produced by the International Standards Organization (ISO) [ISO, 2002] and more specific at the Information Security Services.

5. INFORMATION SECURITY SERVICES

If we want to enforce security on information, we must be able to 'measure' all actions on the information. This 'measurement' can be done in five steps [Von Solms & Eloff, 1999], also called the five pillars of

Information Security. The five pillars are: Identification and Authentication, Authorisation, Confidentiality, Integrity, and Non-denial.

1. Identification and Authentication is the first step towards enforcing information security. First a person that wants to perform a transaction must be identified and during this identification process, this person must also be authenticated, to ensure that somebody else did not provide the claimed identification.
2. Authorisation (also called logical Access Control) is the second step towards enforcing information security. After a person has been identified and authenticated, one must check whether this person has the access right to the requested resource for example, a transaction, a program, a file, or a database.
3. Confidentiality is the third step towards enforcing information security. The assurance that only authorized people may view the contents of the data or software is called protecting the confidentiality of the data or software.
4. Integrity is the fourth step towards enforcing information security. The assurance that only authorized people may change the contents of the data or software is called protecting the integrity of the data or software.
5. Non-denial (or non-repudiability) is the fifth step towards enforcing information security. After changing the contents of the data or software a person must be forced to 'sign' this change so that he cannot at a later stage deny the fact that he made the change.

The next table maps the data mining process onto the five Information Security Services and reveals important security risk areas.

How to read the table:

· The first column of this table explains the data mining process step-by-step by making use of the above-mentioned example.

· The next five columns represent the five Information Security Services.

· Each row, starting with the number one (1), through to seven (7), concentrates on a specific step in the data mining process.

· Each row indicates all the information security risk areas that are present per data mining step by mapping each data mining step to all five Information Security Services. The indication is done by using a tick, for example - √.

· Some of these information security risk areas that are indicated by a number next to the tick is covered in the explanation that follows. The number used in the table, for example - √(5) refers to the fifth explanation of these information security risk areas.

The Data Mining process mapped onto the five Information Security Services

DATA MINING PROCESS (STEPS) FROM THE EXTERNAL TRAVEL AGENT'S OFFICES AND *EXAMPLES SPECIFIC TO THE ILLUSTRATION.*	**INFORMATION SECURITY SERVICES**				
	Identification and Authentication	**Authorisation (Logical Access control)**	**Confidentiality**	**Integrity**	**Non-denial**
1. Define the business problem. Express this in business terms. ***Example****: Structure a Business Class flight-package for the Management of the top business enterprises.*	√		**√(1)**		
2. Build a mining database in other words: Identify a dataset that answers the business question.		**√(2)**	√	√	√

DATA MINING PROCESS (STEPS) FROM THE EXTERNAL TRAVEL AGENT'S OFFICES AND *EXAMPLES SPECIFIC TO THE ILLUSTRATION.*	INFORMATION SECURITY SERVICES				
	Identification and Authentication	**Authorisation (Logical Access control)**	**Confidentiality**	**Integrity**	**Non-denial**
Example: The names of the top business enterprises, the names of their top and senior management as well as the senior public relations staff, each name's current location, nearest airport and travel-record.					
3. Explore the data, in other words: Make some preliminary investigations whether the dataset answers the business question.			√	√**(3)**	

DATA MINING PROCESS (STEPS) FROM THE EXTERNAL TRAVEL AGENT'S OFFICES AND *EXAMPLES SPECIFIC TO THE ILLUSTRATION.*	**INFORMATION SECURITY SERVICES**				
	Identification and Authentication	**Authorisation (Logical Access control)**	**Confidentiality**	**Integrity**	**Non-denial**
***Example**: Will it be possible to build up a history of a specific business enterprise's flight habits and routines with the above-mentioned facts?*					
4. Prepare the data in a format as required by the data mining technique. ***Example**: De normalization or shifting to another platform.*			√	√	√**(4)**
5. Mine the data. ***Example**: Use various data mining*	√**(5)**	√	√	√	

DATA MINING PROCESS (STEPS) FROM THE EXTERNAL TRAVEL AGENT'S OFFICES AND *EXAMPLES SPECIFIC TO THE ILLUSTRATION.*	INFORMATION SECURITY SERVICES				
	Identification and Authentication	**Authorisation (Logical Access control)**	**Confidentiality**	**Integrity**	**Non-denial**
techniques to explain the trends in the dataset.					
6. Do the pattern evaluation in other words: Check whether the business question is answered. ***Example**: Do we have the evidence for an answer as raised? Do we have statistical convincing evidence that a specific business enterprise functions in a specific manner?*			**√(6)**	√	√

DATA MINING PROCESS (STEPS) FROM THE EXTERNAL TRAVEL AGENT'S OFFICES AND *EXAMPLES SPECIFIC TO THE ILLUSTRATION.*	**INFORMATION SECURITY SERVICES**				
	Identification and Authentication	**Authorisation (Logical Access control)**	**Confidentiality**	**Integrity**	**Non-denial**
7. Knowledge presentation in other words: Explain the question and answer in simple terms. ***Example****: Business enterprise X's management tends to travel with Business Class tickets twice a year, once during August and once during March, with a tendency to add their senior public relations staff, also Business Class tickets, every now and then.*			**√(7)**	√	√

Some security risk areas, which are revealed by data mining, will now be explained for the above-mentioned example:

1. The confidentiality of the fact that a data mining exercise is going to take place as well as the outcome of such an exercise must be protected. The assurance that only authorized third party people may view the contents of the data is important when talking about information exposure risks.
2. When building the mining database one must follow a few 'data changing and maintenance' steps (for example data collection, data description, data selection, data cleaning, data integration, metadata construction, load mining database and maintain mining database). The people that have to change and maintain this data must be authorized to do so. They must have the access right to the requested data.
3. When working on this step one will not change the data but one might decide to change the collection of facts to be able to answer the correct business question. This also needs assurance that only authorized people will change the contents of the facts collection and by doing so still protect the integrity of the data.
4. After preparing this data in the required format, it must be 'signed' so that it is possible to trace the person who has done this preparation. This will eliminate any false accusations and statements with the preparation phase.
5. When starting with the different model building techniques and pattern evaluation techniques, one needs to know who is working with the data.
6. At this stage interesting patterns are being evaluated and compared. Are the right people (only authorized people) doing the evaluation?
7. The question and answer now needs some explanation. Are the right people (only authorized people) looking at the results?

6. CONCLUSION

During the data mining process one works intensively with the data involved. The data mining process also expects and allows certain data changes to be made. These initial steps of the data mining process (namely seeing and changing data) cause an information exposure risk. The data mining process as a whole also causes an information exposure risk by making hidden knowledge known.

By agreeing to the fact that a third party may do data mining on an enterprise's data is a step that must be taken with great care and only if one

knows that all the data mining steps in the data mining process, that poses an information exposure risk when cross referenced with the five information security services, has been taken care of.

This research mapped the seven-step data mining process against the five information security services to be able to concentrate on all the information exposure risk areas that are revealed by doing data mining on an enterprise's data. The rest of this research will concentrate on how to minimize the information exposure risk when doing data mining.

7. REFERENCES

[Berry & Linoff, 1997] Berry Michael J and Linoff Gordon with Linoff Gordon S, Data Mining Techniques: For Marketing, Sales, and Customer Support; Wiley, John & Sons, Incorporated, 1997.

[Clifton, 1998] Clifton Chris, Security issues in data warehousing and data mining: panel discussion, in Database Security XI: Status and prospects edited by T Y Lin and Shelly Qian; Chapman & Hall, 1998.

[Han & Kamber, 2001] Han Jiawei and Kamber Micheline, Data Mining: Concepts and Techniques; Academic Press, 2001.

[Inmon, 1996] Inmon W.H., Building the Data Warehouse; New York: John Wiley & Sons, 1996.

[ISO, 2002] ISO 2002, International Standards Organization, http://www.iso.ch, May 2002

[Mattison, 1997] Mattison Rob, Data Warehousing and Data Mining for Telecommunications, Artech House, Incorporated, 1997.

[Pazzani, 2000] Pazzani Michael J, "Knowledge discovery from data?" article in IEEE Intelligent Systems and their applications: Data Mining II, March/April 2000, pp. 10-13.

[Shi, 2000] Shi Yong, "Data mining" article in The IEBM Handbook of Information Technology in Business edited by Milan Zeleny, 2000, pp. 490–495.

[Van Maanen, 2002] Van Maanen Tom, www.van-maanen.com, April 2002

[Von Solms & Eloff, 1999] Von Solms Sebastiaan H. and Eloff Jan H. P., Information Security; Von Solms & Eloff, 1999.