A case study in knowledge acquisition for logistic cargo distribution data mining framework

Puteri N. E. Nohuddin 1,*, Zuraini Zainol 2, Angela S. H. Lee 3, A. Imran Nordin 1, Zaharin Yusoff 3

1Institute of Visual Informatics, National University of Malaysia 43600 Bangi, Selangor, Malaysia
2Department of Computer Science, Faculty of Science and Defence Technology, National Defence University of Malaysia, Sungai Besi Camp 57000 Kuala Lumpur, Malaysia
3Department of Computing and Information Systems, Sunway University, Sunway University, Malaysia

A R T I C L E  I N F O
Article history:
Received 8 August 2017
Received in revised form 16 October 2017
Accepted 10 November 2017

Keywords:
Knowledge acquisition
Data mining
Knowledge representation

A B S T R A C T
Knowledge acquisition is one of important aspect of Knowledge Discovery in Databases to ensure the correct and interesting knowledge is extracted and represented to the stakeholders and decision makers. The process can undertake using several techniques as such in this study, it is using data mining to extract the knowledge patterns and representing the knowledge described using ontology based representation. In this paper, a data set of Logistic Cargo Distribution is selected for the experiment. The dataset describes the shipment of logistic items for the Malaysian Army.

© 2017 The Authors. Published by IASE. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

Processing raw data into meaningful information and knowledge is crucial in data driven decision making. Quality of data can also be an essential issue. Credible data analysis depends on the quality of data from which it is derived. If the data is suspect, concerns may be raised about the quality of decisions that administrators would make based on that data. Losing trust at this stage of the process could make it difficult to rebuild trust moving forward.

Knowledge engineering can be carried out with methods in Data Mining (DM). The ultimate goal for the project undertaken is to develop a methodology for knowledge acquisition from any form of input source culminating in a knowledge base for the development of intelligent systems. The input source for the moment will essentially be in the form of:

- Text (including transcripts of interviews)
- Tables / vectors / matrices
- Databases

but other sources will be envisaged later on, such as:

- Audio
- Images

The targeted knowledge base representation will essentially be ontology-based, but other forms may be employed if found better suited. Techniques for knowledge acquisition will be any combination of:

- Video Multimedia
- Manual (knowledge engineering)
- Data mining / machine learning / statistical methods
- Natural language processing
- etc.

The project is a very long-term project and it is still at an exploratory stage, where case studies are worked on employing various combinations of techniques applied to an array of applications in various domains, all in an effort to abstract generic methods that can be described in formal ways.

This paper presents the results of one such case study on logistic cargo distribution for the military in the data mining domain, where:

- The input data is in the form of vectors
- Data mining techniques will be employed on the data to produce tables
- Manual knowledge engineering techniques will be used to produce the targeted knowledge base.

The remainder of the paper as follows: Section 2 describes some related topics and research work done by others. Followed by Section 3, which elaborates on data mining process using Logistic distribution data then finally in Section 4 concludes the paper.
2. Background and related work

This section reviews the background and relevant literature, which are as follows:

- Knowledge Discovery and Databases (KDD).
- Knowledge Acquisition in Data Mining.
- Knowledge Extraction and Representation

2.1. Knowledge discovery in databases

The terms Knowledge Discovery in Databases (KDD) and Data Mining (DM) have been used interchangeably to describe the process of extracting useful and meaningful information. KDD is defined as the whole process of discovering useful information and knowledge within data, whereas DM is defined as the tasks within the KDD process where tools and mechanism are used to identify (mine) the knowledge of interest -KDD models or steps.

A new generation of computational theories has been formulated and techniques designed to assist in extracting meaningful information from various data sources. The growth of knowledge discovery in databases has brought many researchers into this expanding field. Fig. 1 shows the evolution timeline of DM. Simple reports were created in the 60s where data was obtained from databases, when strong processing was still unavailable, and thus data was only extracted to meet the needs of solving business problems (Solarte, 2002).

![Data mining timeline](image)

Fig. 1: Data mining timeline

In the 80s, individuals and organisations began to demand for information more frequently and wanted results much faster, thus queries were studied and made more known and popular in information retrieval from databases, and at a deeper level as compared to structured reports. A trademark was made on "Database mining", a naming by the Higher National Certificate, where sophisticated algorithms were developed. Later in the 90s, acquiring data became more and more crucial in most day-to-day businesses. Business users needed to respond to immediate questions, and users also wanted their information fast and accurate at the right time to make decisions. This was when the term "Data Mining" emerged in the database community. For instance, the finance sector used data mining to analyze fluctuations in stock prices based on time series forecasting, and more and more industry realized that there is a need to analyze their data in the database.

DM can be characterized as the investigation of extricating valuable data from huge informational collections or databases. With the help of DM, discovery of patterns and hidden knowledge can be shown easily to help in decision making (Wang et al., 2005). It also can be characterized as a spatial DM that is valuable in removing valuable data from enormous measures of information and is profoundly pertinent to applications in which huge information volumes are included, in this manner surpassing human explanatory capacities on analyzing the data (Haluzova, 2008).

A recent survey carried out by (Kohavi, 2001) expressed that DM serves two objectives: knowledge and prediction. These days, DM in different structures is turning into a noteworthy part of business operations. Practically every business procedure includes some form of DM. In term of transportation, a few researchers have been building up an interesting way to deal with street activity administration and blockage control, observing drivers, street mishap investigation, Pavement Management Data. This is a potential aspect to look into (Rahman et al., 2016).

2.2. Knowledge acquisition, extraction and representation

Once a table with sufficient data is obtained from an earlier process, one should be able to extract knowledge from it and represent it in some accepted formalism.

As an example, Fig. 2 indicates how such a table (on the top right-hand side – in this case produced from a database) can be converted into a knowledge base (in this case manually).

2.3. Knowledge acquisition using data mining

Knowledge acquisition (KA) is an important process in knowledge management and knowledge engineering fields (Jantan et al., 2011). KA can be implemented through several methods such as elicitation, collection, analysis, modeling and validation of knowledge (Akerkar and Saija, 2010). In data mining (DM), the KA method is often used for extracting tacit knowledge. Furthermore, the application of DM and machine learning (ML) would help in resolving the KA problem (Ho et al., 2007). DM is basically one of the components in the KDD process (Fig. 3). In general, KDD consists of five main steps: (i) selection, (ii) pre-processing, (iii) transformation, (iv) data mining, and (v) interpretation or evaluation.

Technically, DM is applied to extract or generate interesting information and patterns using algorithms (Dunham, 2006) from large databases. Such valuable information and patterns may assist top level managers in making decision. In order to produce an intelligent decision system, DM tasks and methods can be applied in KA. Generally, there are two main categories in DM tasks: (i) predictive and (ii) descriptive. According to (Dunham, 2006), the predictive model is applied to predict the class of objects using known results from different datasets.
Predictive modelling is often applied in many application areas, for example, (i) crime investigation - to detect crimes and identify suspects after the crime has taken place, (ii) insurance - vehicle insurance to assign risk of incidents to policy holders from information obtained from policy holders, (iii) healthcare - to predict the potential cost or risk associated with managing a specific patient population, (iv) food microbiology, etc. On the other hand, the description is a process of characterizing the general properties of the data. It also identifies patterns and relationships in data.

Some common data mining tasks and techniques are presented in Table 1. These tasks can be applied individually or they can be combined together to perform more sophisticated processes.

### Table 1: Data mining tasks and techniques adopted from (Han et al., 2011; Mourya and Gupta, 2012)

<table>
<thead>
<tr>
<th>DM Tasks</th>
<th>DM Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Decision Tree Induction, Bayesian Classification, Fuzzy Logic, Support Vector Machines (SVM), Rough Set Approach, Genetic Algorithm (GA), etc.</td>
</tr>
<tr>
<td>Clustering</td>
<td>Partitioning Methods, Hierarchical Methods, Density-based Methods, Grid-based Methods, etc.</td>
</tr>
<tr>
<td>Association rules</td>
<td>Frequent Itemset Mining Methods (e.g., Apriori, FP-Growth)</td>
</tr>
</tbody>
</table>

Frequent Pattern Mining (FPM) is used for finding frequent patterns (such as itemsets, subsequences, or substrutures) in data and text mining. The pattern mining concept was proposed by (Agrawal et al., 1993) to find all association rules that can be extracted from frequent patterns in a data set. Technically, this original concept is applied to discover interesting correlations, frequent patterns, and associations between data items in transactional databases (Qiankun and Bhowmick 2003).  

Association Rules Mining (ARM) is an unsupervised data mining technique that comprises the FPM method. Basically, ARM consists of two main steps: frequent pattern mining and association rule generation (Agrawal and Srikant, 1994).
patterns are patterns that frequently exist in the itemsets, where frequency in the target dataset is not less than a threshold value (user specified). For instance, a set of items such as butter and bread are considered as frequent itemsets if both frequently appear in a database. The uncovered relationships between items can be represented in the form of association rules: \( X \rightarrow Y \), where \( X \) and \( Y \) are two items or attribute-value pairs. The quality of a given rules in terms of strength is often measured by two metric values such as support and confidence (Du, 2010; Tan, 2006). Support determines how often a rule is applicable in a given data set whereas confidence determines how frequently items in \( Y \) appear in a transaction that contains \( X \). However, support and confidence have limitations as support threshold can prevent the interesting rules being found (Du, 2010). Therefore, additional measure called lift is applied to discover the interesting rules. Lift is the ratio of confidence to the percentage of cases containing \( Y \). The representation of support, confidence and lift of an association rule \( X \rightarrow Y \) is presented by Eqs. 1-3:

\[
support, s(X \rightarrow Y) = \frac{s(X \cup Y)}{|T|} \quad (1)  
confidence, c(X \rightarrow Y) = \frac{s(X \cup Y)}{s(X)} \quad (2)  
lift(X \rightarrow Y) = \frac{s(X \cup Y)/s(X)}{s(Y)} \quad (3)
\]

If the resulting value of the lift is greater than or equal to 1 then the association rule is considered strong (Eq. 3). Basically, the association rules are generated in two steps. Firstly, the minimum support is used to set all the frequent items. Secondly, each frequent itemset is used to generate all possible rules form it as well as all rules that do not satisfy the minimum confidence level are then removed.

The association rules technique has been widely applied in many real world applications, for example, customer transaction analysis (Agrawal et al., 1993; Najafabadi et al., 2017), healthcare (Konda et al., 2016; Ordonez, 2006; Svarz et al., 2016), predicting flood areas (Harun et al., 2017), text mining (Altuntas et al., 2015; Zainol et al., 2016), wireless sensor networks in smart homes (Rashid et al., 2013; 2015), monitoring activities of dementia patients (Azam et al., 2012), trend analysis in social network (Nohuddin et al., 2010), web mining (Cooley and Srivastava, 2000; Srivastava et al., 2000), software bug analysis, etc. Over the last few decades, a number of FPM algorithms have been proposed for mining association rules. For example, a study conducted by Nasreen et al. (2014) has listed a number of FPM Algorithms such as Apriori, Rapid Association Rules Mining (RARM), Equivalence Class Transformation (ECLAT), FP-Growth, Associated Sensor Pattern of data stream (ASPMS), etc.

3. Logistic cargo distribution data mining framework

This section describes the Logistic Cargo data set and how it is converted in the discretization and normalization as part of a knowledge acquisition process. Then, knowledge patterns are extracted using Frequent Pattern Mining.

3.1. Data source and preprocessing

In this paper, a data set of Logistic Cargo Distribution is selected for the experiment. The dataset describes the shipment of logistic items for the Malaysian Army. The items include vehicles, medicines, military uniforms, and ammunition and repair parts. The datasets are extracted from the records for 2008 to 2009 to form 2 episodes with 12 time stamps each. Cargo items are sent from a few division logistic headquarters to brigades and then to specific battalions in West and East Malaysia. The location of headquarters, brigades and battalions are the spatial attributes of the dataset. These offices are viewed as being sender and receiver nodes and the shipments as links connecting nodes in the network. Each month would have some 100 records. Table 2 shows that each extracted record has 6 attributes: (i) logistic item, (ii) sender, (iii) sender city, (iv) receiver, (v) receiver city, and (vi) shipment cost. Examples of raw data are shown in Table 3.

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Attribute Type</th>
<th>Attribute Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic item</td>
<td>Nominal</td>
<td>(1 ton truck, Ordnance items, uniform...)</td>
</tr>
<tr>
<td>Sender</td>
<td>Nominal</td>
<td>(RAMD, RSD, Artillery)</td>
</tr>
<tr>
<td>Sender city</td>
<td>Nominal</td>
<td>(Kuala Lumpur, Kuantan...)</td>
</tr>
<tr>
<td>Receiver</td>
<td>Nominal</td>
<td>(RAMD, RSD, Artillery)</td>
</tr>
<tr>
<td>Receiver city</td>
<td>Nominal</td>
<td>(Kuala Lumpur, Kuantan...)</td>
</tr>
<tr>
<td>Shipment cost</td>
<td>Continuous</td>
<td>MYR1-500,000</td>
</tr>
</tbody>
</table>

Table 3: Example of raw logistic cargo data

<table>
<thead>
<tr>
<th>Logistic Item</th>
<th>Sender</th>
<th>Sender City</th>
<th>Receiver</th>
<th>Receiver City</th>
<th>Shipment Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 unit Toyota Hilux Double CAB 2.5</td>
<td>92 ATCK, KL</td>
<td>KL</td>
<td>PPT, Sabah</td>
<td>Tawau</td>
<td>22,097.25</td>
</tr>
<tr>
<td>1 Trak 3 Ton Hicom Handalan GS Kargo 11</td>
<td>MKATM-BLP, KL</td>
<td>KL</td>
<td>MKATB2, Kem Kukusan, Tawau, Sabah</td>
<td>Tawau</td>
<td>27,620.00</td>
</tr>
<tr>
<td>3x Trak 3 Ton Hicom Handalan GS Kargo 11</td>
<td>92 DKP, KL</td>
<td>KL</td>
<td>7RAMD, Kem Kukusan, Tawau, Sabah</td>
<td>Tawau</td>
<td>77,138.64</td>
</tr>
<tr>
<td>2x Trak 3 Ton Hicom Handalan GS Kargo 11</td>
<td>92 DKP, KL</td>
<td>KL</td>
<td>3RAMD, Kem Lok Kawi, Sabah</td>
<td>Lok Kawi</td>
<td>51,425.76</td>
</tr>
</tbody>
</table>

For this study, Frequent Pattern Mining (FPM) and part of Association Rule Mining, are used to extract frequent patterns of cargo items that are frequently sent to the military camps. Discretization and normalization processes are used to convert the input data, presented in some non-binary format.
into the binary valued format. This step is necessary because the data mining techniques to be used for FPM would only operate with binary valued data (0-1 data). Discretization converts the original dataset attributes with continuous data values into \((1, ..., N)\) sub-ranges such that each sub-range is identified by a unique integer label. Normalization converts data attributes with nominal values into unique integer labels/columns. For the experiments in this research, any attribute with continuous data types are divided into 10 sub-ranges and the attributes with integer data types are divided into 5 sub-ranges. Thus, the data format conversion maintains the nature of the data while at the same time permitting the application of FPM algorithms.

Fig. 4 summarizes the discretization and normalization conducted with respect to the Logistic Cargo data. As a result, the attributes in the data set are normalized to 201 attributes.

![Fig. 4: Normalised data attributes and labels](image)

### 3.2. Knowledge pattern interpretation

FPM generates a set of combination patterns that describes the combination of attributes. Three minimum support threshold values of 2%, 3%, and 5% are used in this study. Table 4 shows the number of frequent patterns identified using the Logistic Cargo data.

<table>
<thead>
<tr>
<th>Year</th>
<th>2%</th>
<th>3%</th>
<th>4%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>3491</td>
<td>3491</td>
<td>3491</td>
</tr>
<tr>
<td>2009</td>
<td>2761</td>
<td>2761</td>
<td>2609</td>
</tr>
</tbody>
</table>

Table 5 provides some examples of frequent patterns associated with the distribution of logistic items extracted from the Logistic Cargo data set. The frequent patterns feature the following attributes: logistic item, shipment cost, sender ID, receiver ID and city location of sender and receiver. Again, the identified Logistic Cargo frequent patterns have support values for January to December.

In Table 5, some frequent patterns are presented such as \{Logistic items = Ordnance items\} = \{3, 7, 3, 2, 6, 1, 3, 1, 3, 2, 1\}, which means Ordnance items are amongst the frequent items that have been distributed between January and December. Another example, \{Sender city = Batu Caves, Logistic items = 1 tonne truck\} = \{0, 2, 3, 1, 1, 4, 1, 0, 0, 0, 0\} means that the sender office is from Batu Caves, and have sent two 1 tonne trucks in February, three 1 tonne trucks in March, and so on.

### 3.3. Knowledge acquisition from data mining

For the final step, information generated in the form of tables in the DM process is looked at and converted to knowledge representation. It is currently done manually, but work is in progress to produce a more semi-automated version.

The top of Fig. 5 gives the base contents of a Transaction, namely the Month (time stamp), Item, Location of Sender, Location of Recipient, and Reference Number of the Transaction. The bottom part gives a general Ontology of the components involved, where an organisation may be subdivided into Suppliers and Clients, in reference to the Logistic dataset Supplier is referred as Sender and Client is referred as Receiver, and each of these concepts have their attributes and values (those below would inherit from the concepts above under the IS_A relation). These organisations would possess (Have) Items.

Transactions are a more dynamic concept, which is usually less apparent in many declarative knowledge formalisms (such as Ontologies). Nonetheless, we have declared here what a Transaction should be in Fig. 6. The frequent transactions are recorded as an attribute of the transition concept.

These concepts can naturally be linked to an existing knowledge base containing the entities involved, which would then provide further (or contextual) knowledge. Another clear advantage is that this knowledge base provides a historical record of transactions for future reference, and any further exercise (say for frequent patterns) may be done incrementally, i.e. not having to begin again. As such, the DM exercise has helped refine the knowledge base, while the knowledge base will help improve further DM efforts.
4. Conclusion and future work

This paper presents part of a study towards a goal of developing a methodology for knowledge acquisition from an input source culminating in a knowledge base for the development of intelligent systems. The project is a very long-term project and it is still at an exploratory stage, and the work presented here is a case study on logistic cargo distribution for the military, where the input data is in the form of vectors, and data mining techniques are employed on the data to produce tables, on which manual knowledge engineering techniques are used to produce the targeted knowledge base in the form of an ontology.

References

