

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

ARTICLE INFO

Article history:

Received 8 August 2017

Received in revised form

16 October 2017

Accepted 10 November 2017

Keywords:

Knowledge acquisition

Data mining

Knowledge representation

ABSTRACT

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

- Video Multimedia

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:

- 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 broad 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.

* Corresponding Author.

Email Address: puteri.ivi@ukm.edu.my (P. N. E. Nohuddin)

<https://doi.org/10.21833/ijaas.2018.01.002>

2313-626X/© 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/>)

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, and
- 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).

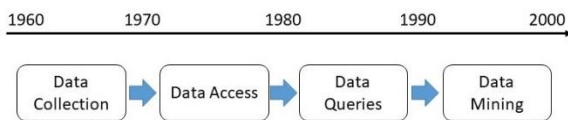


Fig. 1: Data mining timeline

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 Sajja, 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.

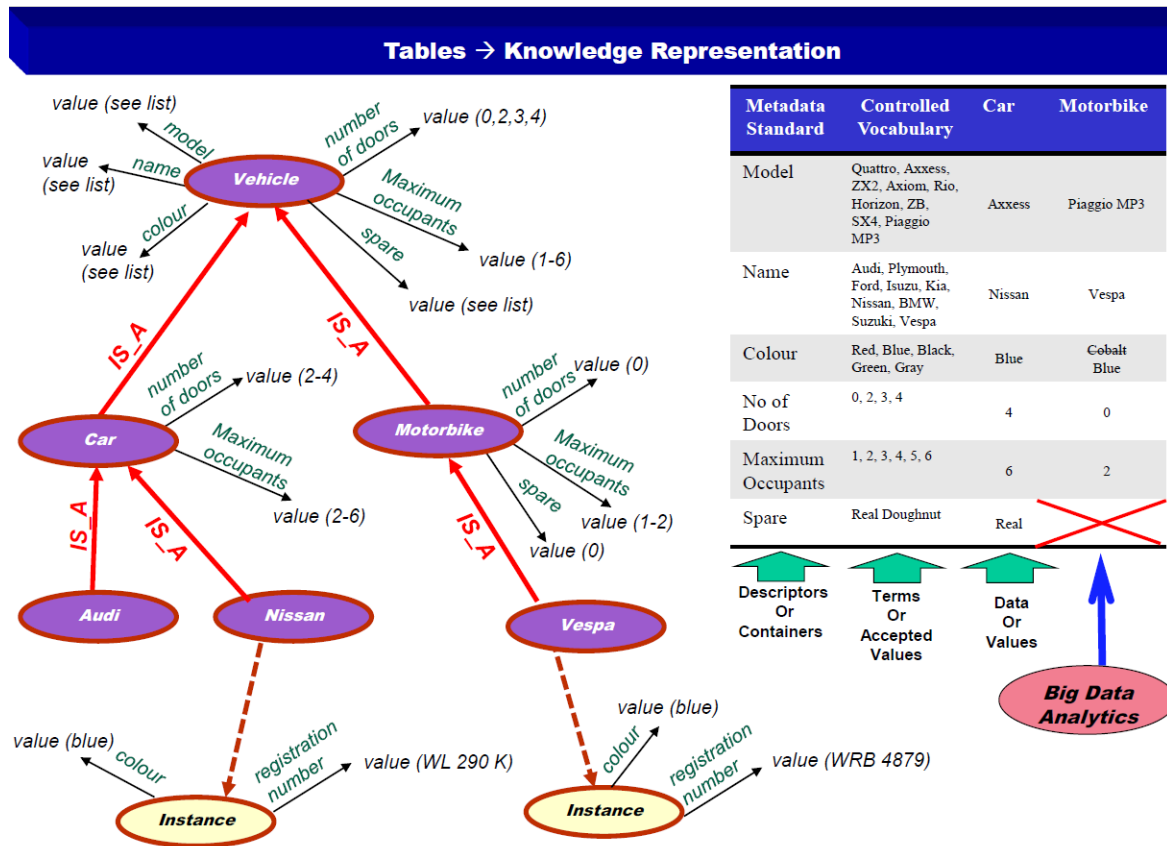


Fig. 2: Tables to knowledge representation

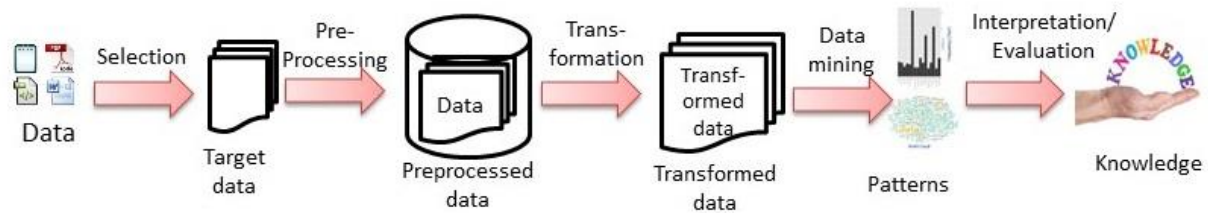


Fig. 3: KDD Process (Fayyad et al., 1996)

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)

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

Frequent Pattern Mining (FPM) is used for finding frequent patterns (such as itemsets, subsequences, or substructures) 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). Frequent

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 term 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:

$$\text{support}, s(X \rightarrow Y) = s(X \cup Y) / |T| \quad (1)$$

$$\text{confidence}, c(X \rightarrow Y) = s(X \cup Y) / (X) \quad (2)$$

$$\text{lift}(X \rightarrow Y) = s(X \cup Y) / s(X) \cdot s(Y) \quad (3)$$

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 from 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; Ordóñez, 2006; Szvarça 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 2: List of attributes

Attribute Name	Attribute Type	Attribute Value
Logistic item	Nominal	{1 ton truck, Ordnance items, uniform...}
Sender	Nominal	{RAMD, RSD, Artillery}
Sender city	Nominal	{Kuala Lumpur, Kuantan...}
Receiver	Nominal	{RAMD, RSD, Artillery}
Receiver city	Nominal	{Kuala Lumpur, Kuantan...}
Shipment cost	Continuous	MYR1-500,000

Table 3: Example of raw logistic cargo data

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

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.

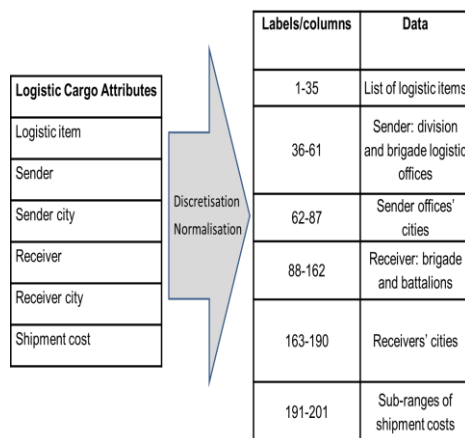


Fig. 4: Normalised data attributes and labels

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 4: Number of frequent pattern trends using support thresholds 2%, 3% and 5%

Year	2%	3%	4%
2008	3491	3491	3491
2009	2761	2761	2609

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, 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, 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.

Table 5: Examples of frequent patterns

No	Frequent Patterns	Support count for 12 months
1.	{Logistic items = Ordnance items}	{3, 7, 3, 2, 6, 1, 3, 1, 3, 3, 2, 1}
2.	{Sender city = Batu Caves, Logistic items = 1 tonne truck}	{0, 2, 3, 1, 1, 4, 1, 0, 0, 0, 0, 0}
3.	{0, 0, 2, 1, 0, 0, 1, 0, 0, 0, 1, 0}	{0, 11, 8, 4, 1, 8, 3, 0, 3, 1, 3, 0}
4.	{MYR50001 <= cost <= MYR100000, Receiver city = Sibiu, Sender city = Batu Kentonmen, Sender = 91 DPO, Logistic items = Ordnance items}	

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.

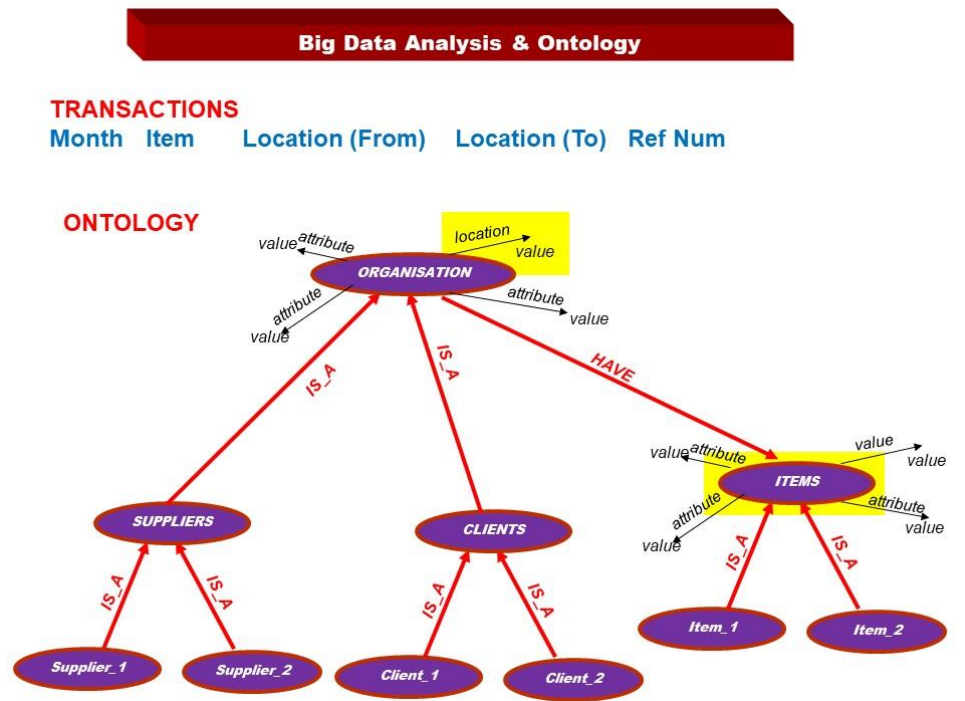


Fig. 5: Big data analysis and ontology

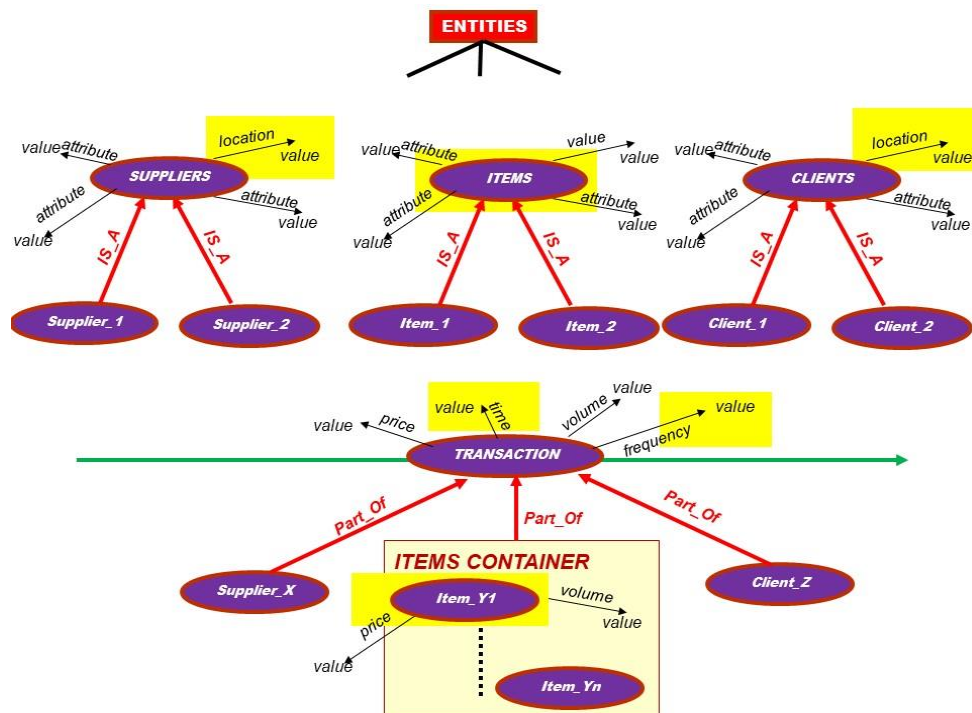


Fig. 6: Knowledge acquired

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

- Agrawal R and Srikant R (1994). Fast algorithms for mining association rules. In the 20th International Conference on Very Large Data Bases, VLDB, 1215: 487-499.
- Agrawal R, Imieliński T, and Swami A (1993). Mining association rules between sets of items in large databases. In the ACM SIGMOD International Conference on Management of Data,

- Washington, D.C., USA, 22(2): 207-216. <https://doi.org/10.1145/170036.170072>
- Akerkar R and Sajja P (2010). Knowledge-based systems. Jones and Bartlett Publishers, Burlington, Massachusetts, USA.
- Altuntas S, Dereli T, and Kusiak A (2015). Analysis of patent documents with weighted association rules. *Technological Forecasting and Social Change*, 92: 249-262.
- Azam M, Loo J, Naeem U, Khan S, Lasebae A, and Gemikonakli O (2012). A framework to recognise daily life activities with wireless proximity and object usage data. In the 23rd IEEE International Symposium on Personal, Indoor and Mobile Radio Communication, IEEE, Sydney, Australia: 590-595.
- Cooley RW and Srivastava J (2000). Web usage mining: Discovery and application of interesting patterns from web data. University of Minnesota, Minneapolis, USA.
- Du H (2010). Data mining techniques and applications: An introduction. Cengage Learning, Boston, Massachusetts, USA.
- Dunham MH (2006). Data mining: Introductory and advanced topics. Pearson Education India, Bengaluru, India.
- Fayyad U, Piatetsky-Shapiro G, and Smyth P (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11): 27-34.
- Haluzova P (2008). Effective data mining for a transportation information system. *Acta Polytechnica*, 48(1): 24-29.
- Han J, Pei J, and Kamber M (2011). Data mining: Concepts and techniques. Elsevier, Amsterdam, Netherlands.
- Harun NA, Makhtar M, Aziz AA, Zakaria ZA, Abdullah FS, and Jusoh JA (2017). The application of apriori algorithm in predicting flood areas. *International Journal on Advanced Science, Engineering and Information Technology*, 7(3): 763-769.
- Ho TB, Kawasaki S, and Granat J (2007). Knowledge acquisition by machine learning and data mining. In: Wierzbicki AP and Nakamori Y (Eds.), *Creative environments: Issues of creativity support for the knowledge civilization age*: 69-91. Springer Berlin Heidelberg, Berlin, Germany.
- Jantan H, Hamdan AR, and Othman ZA (2011). Talent knowledge acquisition using data mining classification techniques. In the 3rd Conference on Data Mining and Optimization, IEEE, Putrajaya, Malaysia: 32-37. <https://doi.org/10.1109/DMO.2011.5976501>
- Kohavi R (2001). Data mining and visualization. In the 6th Annual Symposium on Frontiers of Engineering, National Academy Press, Washington, D.C., USA: 30-40.
- Konda S, Balmuri KR, Basireddy RR, and Mogili R (2016). Hybrid approach for prediction of cardiovascular disease using class association rules and MLP. *International Journal of Electrical and Computer Engineering*, 6(4): 1800-1810.
- Mourya S and Gupta S (2012). Data mining and data warehousing. Alpha Science International, Oxford, UK.
- Najafabadi MK, Mahrin MNR, Chuprat S, and Sarkan HM (2017). Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data. *Computers in Human Behavior*, 67: 113-128.
- Nasreen S, Azam MA, Shehzad K, Naeem U, and Ghazanfar MA (2014). Frequent pattern mining algorithms for finding associated frequent patterns for data streams: A survey. *Procedia Computer Science*, 37: 109-116.
- Nohuddin PN, Christley R, Coenen F, Patel Y, Setzkorn C, and Williams S (2010). Frequent pattern trend analysis in social networks. In the International Conference on Advanced Data Mining and Applications, Springer, Berlin, Heidelberg, 358-369.
- Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine*, 10(2): 334-343.
- Qiankun Z and Bhowmick SS (2003). Association rule mining: A survey. Technical Report, CAIS, Nanyang Technological University, Singapore, India.
- Rahman MF, Shamsuddin SM, Hassan S and Abu Haris N (2016). A review of KDD-Data mining framework and its application in logistics and transportation. *International Journal of Supply Chain Management*, 5(2): 77-84.
- Rashid MM, Gondal I, and Kamruzzaman J (2013). Mining associated sensor patterns for data stream of wireless sensor networks. In the 8th ACM workshop on Performance Monitoring and Measurement of Heterogeneous Wireless and Wired Networks, ACM, Barcelona, Spain: 91-98. <https://doi.org/10.1145/2512840.2512853>
- Rashid MM, Gondal I, and Kamruzzaman J (2015). Mining associated patterns from wireless sensor networks. *IEEE Transactions on Computers*, 64(7): 1998-2011.
- Solarte J (2002). A proposed data mining methodology and its application to industrial engineering. M.Sc. Thesis, University of Tennessee, Knoxville, Tennessee, USA.
- Srivastava J, Cooley R, Deshpande M, and Tan PN (2000). Web usage mining: Discovery and applications of usage patterns from web data. *Acm Sigkdd Explorations Newsletter*, 1(2): 12-23.
- Szvarca RR, Ioshii SO, Carvalho DR, and Sokoloski WF (2016). Temporal association rules in breast cancer. *Iberoamerican Journal of Applied Computing*, 4(3): 14-20.
- Tan PN (2006). Introduction to data mining. Pearson Education India, Bengaluru, India.
- Wang W, Chen H, and Bell MC (2005). Vehicle breakdown duration modelling. *Journal of Transportation and Statistics*, 8(1): 75-84.
- Zainol Z, Nohuddin PN, Jaymes MTH, and Marzukhi S (2016). Discovering "interesting" keyword patterns in Hadith chapter documents. In the International Conference on Information and Communication Technology, IEEE, Kuala Lumpur, Malaysia: 104-108. <https://doi.org/10.1109/ICICTM.2016.7890785>