Knowledge Discovery Database (KDD)-Data Mining Application in Transportation

Fauziah Abdul Rahman
Faculty of Technical Foundation,
Universiti Kuala Lumpur
Malaysian Institute of Industrial Technology (MITEC)
fauziahar@mitec.unikl.edu.my

Mohammad Ishak Desa
Faculty of Computing
Universiti Teknologi Malaysia
Johor, Malaysia
mishak@utm.my

Antoni Wibowo
Faculty of Computing
Universiti Teknologi Malaysia
Johor, Malaysia
antoni@utm.my

Norhaidah Abu Haris
Faculty of Software Engineering
Universiti Kuala Lumpur,
Malaysia
norhaidah@unikl.edu.my

Abstract— ***In this paper, an understanding and a review of data mining (DM) development and its applications in logistics and specifically transportation are highlighted. Even though data mining has been successful in becoming a major component of various business processes and applications, the benefits and real-world expectations are very important to consider. It is also surprising to note that very little is known to date about the usefulness of applying data mining in transport related research. From the literature, the frameworks for carrying out knowledge discovery and data mining have been revised over the years to meet the business expectations. In this paper, we apply CRISP-DM for formulating effective tire maintenance strategy within the context of a Malaysian's logistics company. The results of applying CRISP-DM for tire maintenance decisions are presented and discussed.***

Keywords— ***Data Mining; Knowledge Discovery Database-Data Mining (KDD-DM); Domain-Driven Data Mining (DDDM); Actionable Knowledge Discovery (AKD); Logistics and Transportation***

I. Introduction

Logistics can be understood as a subset of supply chain management performance. Logistics can be defined as strategically managing the procurement, involve the movement of materials as well as storage of materials, parts and finished products inventory and related with the information flows, through the organization at a maximum profits with minimum costs in fulfillment of orders[3].American Council Logistics Management also defined logistics as the process of planning, implementing and controlling the efficient, cost effective flow and storage of raw materials, in-process inventory, finished goods and related information from point of origin to point of consumption for the purpose of conforming to customer's requirements. Logistics became as a planning orientation and framework that seek to create a single plan for the flow of product and information through a business. Organization deals with supplier and customer to measures the quantity of material that passes through a given network per unit of time. Thus to achieve the logistics objectives, it builds upon the logistics framework to achieve the linkage within particular organization and with the processes of other organization. The incessant economic and industrial activities around the globe and the splurge of exports and imports continue to impose greater demands on shipping and cargo industry. Hence, the traditional transportation vendors not only strive to deliver cargo securely and accurately to customers on time but also consider reducing cost and flexibility dispatching vehicles as well as staff [2]. Thus, in order reducing costs at time related positioning resources, logistics scheduling problem has gained increasing importance with the development of supply chain management. Logistics scheduling has to deal with job delivery and transportation issues [4]. This includes minimize the sum of weight job delivery and the total transportation cost. The worldwide cost to industry of outsourced logistics problems in 1995 have been highlighted in three (3) areas, maintenance of fleets, distribution and delivery stock. It is estimated to be AUS$ 900 Billion dollars of which the cost to Australian organizations is two billion.

II. LOGISTICS AND FLEET MAINTENANCE PROBLEMS

Given the proliferation and complexity of some logistics problems, the application of computer systems, particularly decision support system is expected to increase significantly. However, current software tools for decision support in logistics do not totally address the combination of these characteristics. Software developed specifically for logistics problems are usually offer the reporting and tracking variety and or automate routine tasks but offer little support for decision making. Tools currently available for decision support are usually problem specific or too general to complement the decision process in logistics problems. There were four characteristics that have been identified as the most real world logistics problem. This includes large decision space; consists of a large of number of decision variables and

ISBN No. 978-602-70504-0-2 (Print)
ISBN No. 978-602-70504-1-9 (media)

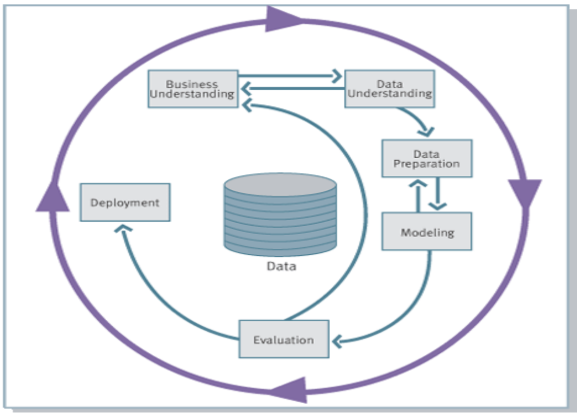
possible options or strategies, availability of real time data; business modern that have extensive data collection capabilities that provides operational data which can be used for effective optimization and communication in real time which uncertainty problems in making decision because of the uncertain and incomplete knowledge about future circumstances, numerous decision makers and complexity and dynamic behavior among the logistics components and interconnectivity. A general decision support system framework for logistics has been done and a mapping between research areas and logistics problem characteristics has been highlighted [11]. Three (3) outsourced logistics problems identified. It was fleet maintenance, distribution and delivery stock which are related to uncertainty research area; thus it is proved that those areas are among the most relevant to logistics problem in a real business world.

III. PROBLEM'S BACKGROUND

Previously, there are many researches related to transportation involving vehicle routing, vehicle scheduling, fleet preventive maintenance related with time windows in job delivery and transportations using statistical method. Using statistical method, sometimes one can find patterns are not significant in reality. Data mining is a legitimate activity as long as one understands how to do it correctly. But very little is known to date about the usefulness of applying data mining in logistics and transport related research. Nowadays, the computer based systems are being used to automatically diagnose problems in vehicles in order to overcome some of the disadvantages associated with relying completely on experienced personnel. Typically, a computer based system utilizes a mapping between the observed symptoms of the failure and the equipment problems using techniques such as table look-ups, a symptom problem matrices and production rules. These techniques work well for simplified systems having simple mappings between the symptoms and problems. However, complex equipment and process diagnostics seldom have simple correspondences. In addition not all symptoms are necessarily present if problem has occurred, thus making other approaches more cumbersome [21]. These approaches either take a considerable amount of time before a failure are diagnosed or provides less than reliable results, or are unable to work well in complex systems. There is a need to be able to quickly and efficiently determine the cause of failures occurring in the vehicle maintenance system, while minimizing the need of human intervention [21].Having a direct access to systems data from remote vehicles would helpful in optimizing vehicle maintenance scheduling, route planning and minimize downtime from unexpected breakdown such as track vehicles with artificial intelligence but depending on it alone was costly [21]. The research also shown that, the existence fleet management can only analyze records after incident-occurrence and cannot analyze vehicle status in a real time. Even though the future system can be integrated with real time technology such as Global Positioning System (GPS), that can provide more valuable information, it will lead to data accumulation [6]. Thus in identifying imminent system failures or failure prognostics, better diagnostics data in the system is another way to help in enhancing the capability of maintainers at minimal cost where data mining is applied.

IV. CRISP-KDD DM METHODOLOGY

The whole process is sometimes called as knowledge discovery databases (KDD). This was the first generation of KDD where DM process attached together in the KDD life cycle to ensure a discover knowledge can meets the business requirements. Nowadays researchers with strong industrial engagement realized the need from DM to KDD to deliver useful knowledge for the business decision making. Traditionally, one standard, named CRISP-DM (Cross-Industry Standard Process for Data Mining) Methodology, determine the process step helps to avoid common mistakes[7][23]. It is important to understand each phases before implementing DM process.



a

Figure 1. Cross-Industry Standard Process for Data Mining Methodology (CRISP-DM))

The first phase is business understanding where to understand what is really to be accomplished. This task involves more detailed fact-finding about all the resources, assumptions and other factors that should consider in determining the data analysis goal. Second phase is data understanding that investigates a variety of descriptive data characteristics (count of entities in table, frequency of attribute value, average values and etc.). Third phase is data preparation which is the most difficult and time-consuming element in KDD process. The goal is to choose relevant data from available data, and to represent it in a form which is suitable for the analytical methods that are applied. Data preparation includes activities like data selection, filtering, transformation, creation, integration and formatting. The fourth phase is modeling which is the use of analytical methods (algorithms). There are many different methods and most suitable one must be chosen. This phase is also verifying the quality of the model such as testing in the independent data matrix, cross validation and others. The fifth phase is evaluation where the interpretation and evaluation of the discovered knowledge. In a decade, CRISP-DM life cycle representation of DM process seems to become more dominant [16]. However using this

ISBN No. 978-602-70504-0-2 (Print)
ISBN No. 978-602-70504-1-9 (media)

traditional framework represented some issues when the deployment stages are taken. The framework life cycle is sequential and linear. Even though the feedback loops are mentioned the sequential, natures of the representation suggest an ordering of the knowledge space and its exploration is not appropriately characterized the hierarchical and interactive network features of corporate knowledge space or can be as dynamic of DM [16]. CRISP-DM is a data centered-heavily depend on data itself [25] or data methodology or called as Data-oriented base framework. Current dominant situations are narrow focus and over emphasized by innovative data-driven and algorithm-driven research. In the real world scenarios, challenges always come from specific domain problems which back to the goal of DM towards business concerns, hence the objectives and goals of applying KDD are basically problem solving to satisfy real user needs.

V. CASE STUDY

A. ASL Company

ASL is one of the logistics companies in Malaysia. The main operation activities of the land transportation involving tankers and cargo trailers are transportation of palm oil, dry cargo, palm fruit, latex and courier. These transportation vehicles are the most contributing costs of operations and maintenance. However, tire is the major contributing costs beside fuel but it can be rethreaded and help in reducing operating cost. Due to this problem, the company has implemented an application called Tire Management System (TMS). Currently, TMS is used by the company to produce reports for tire maintenance planning and operational decisions. It is however, analyzing such enormous data using conventional technique is mind boggling task for the company. The current implementation of TMS requires the administrator from all depots to key-in data on tire maintenance into the system that are given by the duty drivers and hence causing inaccurate data and time consuming. Every depot (branch) has one workshop headed by a Senior Foreman and supported by staff consists of mechanics, welders and craftmens. However, it is imperative that companies such as delivery drivers know what is going on with their vehicle’s tire at all times not only depends on the routine maintenance services. In this study KDD-DM is applied to the available TMS to gain useful knowledge for effective tire maintenance decision making.

B. CRISP-KDD Stages

Stage 1: Business understanding related with goals of analysis
Tire was identified to be among the most contributing cost to ASL. From the year 2008 until now, it is reported that tire maintenance cost is the second largest variable cost and needed to be rectified after diesel. Since the tire cost is not available in TMS system, we can only analyze possible knowledge gain from the analysis. Therefore, we can correlate it with the tire cost in the future. The objective of the analysis is to reduce the tire cost by identifying the most contribution tire cost by detached reasons (DR) and the range of its journey as well as by tire attached position based on the standard tire configuration by ASL. The knowledge obtained from the journey of the trucks and the detached reasons causes is valuable for other knowledge; for instance, we can correlate it with the tire position that shows that particular tire position need to be inspected more frequently. In current practice, ASL try to decrease the used of new tire and the expensive tire brand.

Stage 2: Business understanding related with goals of analysis. This stage involves the process of collecting data and investigates a variety of descriptive data characteristics for instance, count of entities in table, frequency of attribute value and average values using Exploratory Data Analysis (EDA). The available data come from TMS system of ASL where all the tire transactions of three (3) main depots have been recorded.

Stage 3: Data preparation
The goal is to choose relevant data from available data, and to represent it in a form which is suitable for the analytical methods that are applied. This stage involves DC activities which refer to the CRISP-DM KDD framework as discussed in Table 1. Several methods has been used including missing value method in statistical analysis tool, parsing method where detection of lexical errors (syntactical error) and domain errors of records for instance, eliminated or duplicated negative values, integrity constraint enforcement method by adding updates existing records for instance tire brand based on objective of analysis and also did data transformation where normalization and standardization of records into uniform format have been done. Additionally, the inconsistencies of data also have been removed. It was based on domain experts. For example, for "botak rata", it shouldn't be less than 60, 000 km journey. The domain knowledge was gained from an interview conducted with domain experts and domain users. According to [5] high quality data that has been clean needs to pass a set of quality criteria as in Figure 4. After DC process, only 1016 records with four (4) variables will be used in modeling stage.

Stage 4: Modelling
Previous studies hows that there are several different techniques may be used for the same data mining problem. For the case study, the Classification method using C4.5 Decision tree technique was chosen because it produce set of rules that easy to understand by human [6].

Stage 5: Evaluation
The result indicated that the most contributing cost of DR are "Meletup" and "Nampak Steel Belt". Researchers found out that Tire Attached Position (AP) for RR3 and TL3 contributed the most tire cost because of the both DR. For TL3 position, it is believe that a new tire has been attached to this position. Unfortunately, the tire life span for the new tire attached was ended between journey range within 58,000km until 132,000 km journey. It was surprised that the result shown that D.R for "Nampak steel belt" and "Meletup" occurred within the range

ISBN No. 978-602-70504-0-2 (Print)
ISBN No. 978-602-70504-1-9 (media)

of journey >=58,729km until less than <132,692 km. Based on the domain expertise and domain knowledge, a new tire always can be use until 80,000km with D.R "Botak rata" which cause the dye tire is more cheaper than a new tire. It is believed that a new tire will be longer life span if it is meet the routine inspection. ASL company has the policy on attachment of tire based on the truck's axles. In this case study the data was based on there (3) axles. It was surprisingly found out that most of the AP dye tires that were supposedly rotate using a dye tire such as position RL3, BL1.BR4,BL3,BR1,BL4,BR3 and BL2 were replace with new tire where contributed to the tire cost. The rules expected did not produce and met the objectives of the analysis. It can only determine the frequency of the tire usage based on AP and D.R based on journey per kilometer.

Stage 6: Deployment
The researchers found that the results cannot be deployed and it is noted that the current classification performance is inaccurate. It is necessary to loop back to the data preparation phase until the classification performance is increase.

C. Result and Discussion

As a conclusion, researchers found that the results cannot be deployed and it is noted that the current classification performance is inaccurate. The previous studies shows that the researchers need to explore other DM techniques rather than Classification Decision Tree C4.5 technique to achieve the objective of the analysis. Some others DM methods used and discussed in previous researchers including an Association Rules and Clustering techniques including optimization areas to have optimum decision making. The DC process were done based on the DC process in the previous study [4][5][7]. In the real world scenarios, domain experts are slightly important for data validation in CRISP-KDD methodology. Researchers have difficulty experienced in doing the existing DC process in term of long time DC process that produced an inaccurate result. A formalize DC process that generate high data quality are critically needed for the organization specifically for ASL as one of the logistics company in Malaysia.

ACKNOWLEDGMENT

We would like to express our thanks specifically to ASL company for their support of domain knowledge and data extraction and to Universiti Kuala Lumpur (UniKL) for the financial support to the first author to undertake a PhD research at Universiti Teknologi Malaysia (UTM).

REFERENCES

[1] Dan Luo, Longbing Cao, Chao Luo, Chengqi Zhang and Weiyuan Wang, 2008. Towards Business Interestingness in Actionable Knowledge Discovery. *Proceeding of the 2008 conference on Applications of Data Mining in E-Business and Finance*(IOS Press Amsterdam, The Netherlands, The Netherlands, 2008). DOI= http://dl.acm.org/citation.cfm?id=1565639.1565650

[2] Cao Longbing and Zhang Chengqi. 2007. The Evolution of KDD: Towards Domain-Driven Data Mining, . *International Journal of Pattern Recognition and Artificial Intelligence, 21(4), 677-692.*

[3] Longbing Cao, 2008. Domain Driven Data Mining: Challenges and Prospects. *Journal on Knowledge and Data Engineering.* 6 (22)(June 2010), 755-769.

[4] Kalaivany Natarajan, Jiuyong Li and Andy Koronios,2009. Data Mining Techniques or Data Cleaning. *In Proceedings of the 4th World Congress on Engineering Asset Management* (Athens, Greece, 28 – 30, September). DOI= http://dx.doi.org/10.1007/978-0-85729-320-6_91

[5] R.KAVITHA KUMAR and et.el . 2011.Attribute Correction-Data Cleaning using Association Rule and Cluestering Method. *International Journal of Data Mining & Knowledge Management Process (IJDKP.* 1(2)(March 2011), 22-32.

[6] Sang Jun Lee and et. el. 2001. A Review of Data Mining Technique, *Industrial Management & Data Systems,*101 (1).41-46.

[7] Daniel T.Larose. 2005. Discovering Knowledge in Data:An introduction on Data Mining. Book. 27-65.

[8] Hasimah Hj Mohamed and et. el. 2011. E-Clean: A Data Cleaning Framework for Patient Data. 2011. *First International Conference on Informatics and Computational Intelligence.* DOI= http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=6141651\

[9] Erhard Rahm, Hong Hai Do. 2000. Data Cleaning: Problems and Current Approaches. *IEEE Data(base) Engineering Bulletin - DEBU Journa*l,. 23(4). 3-13.

[10] Lior Rokach and et. el. 2010. Data Mining with Decision Trees:Theory and Applications. World Scientific Publishing Co. Pte. Ltd. 1-214.

[11] Paolo Giudici et. el.Applied Data Mining for Business and Industry. 2010. Johd Wiley & Sons Ltd. 219-225.

[12] Alex A. Freitas. Data Mining and Knowledge Discovery with Evolutionary Algorithms. 2004. Springer. 65-75.

[13] Tamraparni Dasu and et. el. Explorattory Data Mining and Data Cleaning. 2004. A John Wiley &Sons, Inc., Publication. 99-189.

[14] Tsau Youn Lin and et. el.. Direct Data Mining of Rules from Data with Missing Values. 2005. Springer.233-264.

[15] Longbing Cao.2008. Introduction to Domain Driven Data Mining. Data Mining For Business Application. Springer 2009.

[16] Longbing Cao and Chengqi Zhang.2007. The Evolution of KDD: Towards Domain-Driven Data Mining. International Journal of Pattern Recognition and Artificial Intelligence. Vol. 21. No. 4 (2007). Pages 677-692.

[17] Lily Sun, Cleopa John Mushi.2010. Case-based analysis in user requirements modeling for knowledge construction. Elsevier.

[18] P. Haluzova, 2008. Effective Data Mining for a Transportation Information Systems. Czech Technical University Publishing

[19] Rayid Ghani and Carlos Soares . 2006. KDD-2006 Workshop.Accenture Technology Labs.

[20]Sudhir Kumar Barai. 2003. Data Mining Applications in Transportation Engineering. India Institute of Technology Kharagpur, India.

[21] U.S Pattern Storm.2001.Vehicle Maintenance Management System and Method.

[22] Thomas Young and et.al.2010.Utilizing Data Mining to Influence Maintenance Actions.

[23] WenQun Wang, Haibo and Magaret. 2004. Vehicle Breakdown Duration Modeling. Journal of Transportation and Statistics.

[24] William R. King, Peter V. Marks, Jr., and Scott McCoy. 2002. The Most Important Issues in Knowledge Management. Communications Of the ACM, September 2002 Vol. 45 No 49.

[25] Zhengxiang and et.al.2009. Research on Domain-Driven Actionable Knowledge Discovery. Springer 2009.

ISBN No. 978-602-70504-0-2 (Print)
ISBN No. 978-602-70504-1-9 (media)