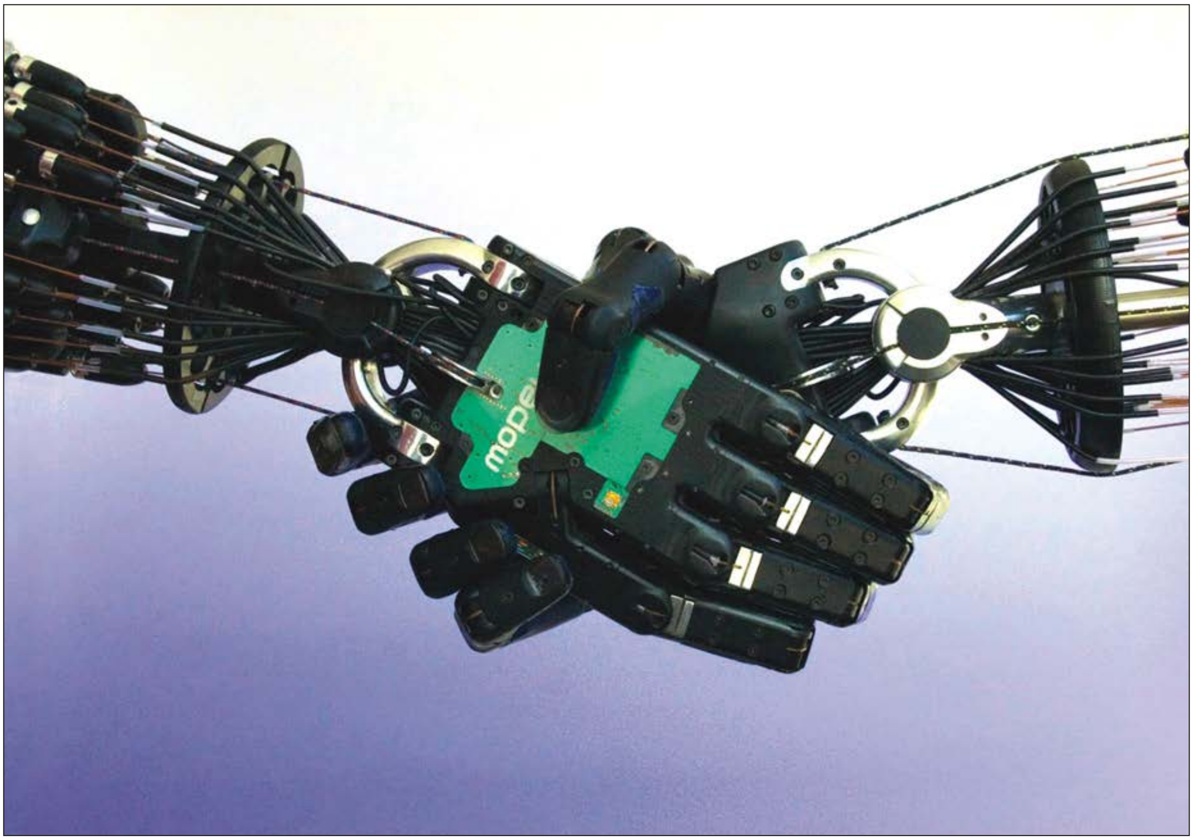


Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Artificial Intelligence

ISSN 1994-5450
DOI: 10.3923/jai.2016.72.77

Research Article
An Improvement of Knowledge Discovery Database (KDD) Framework for Effective Decision

[1]Fauziah Abdul Rahman, [2]Muhammad Ishak Desa, [2]Antoni Wibowo and [4]Norhaidah A. Haris

[1]Malaysian Institute of Industrial Technology (MITEC), Universiti Kuala Lumpur, Persiaran Sinaran Ilmu, Bandar Seri Alam, Masai, 81750 Johor, Malaysia
[2]Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
[4]Malaysian of Information Technology (MIIT), Universiti Kuala Lumpur, 1016, Jalan Sultan Ismail, 50250 Kuala Lumpur, Malaysia

Abstract

In this study, an understanding and a review of Knowledge Discovery Database (KDD) development and its applications in tire maintenance 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 knowledge discovery 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 requirements. The Domain Driven Data Mining (DDDM) is one of the KDD frameworks often used for this purpose. In this study, we apply DDDM-KDD for formulating effective tire maintenance strategy within the context of a Malaysian's logistics company. We also discussed the weaknesses of the results from DDDM-KDD and emphasize the important of using the next generation of KDD framework Actionable Knowledge Discovery (AKD) for an effective decision. The direction flow of research, research methods use and contribution of research also are highlighted.

Key words: KDD, AKD, data mining, ASL, data cleaning study

Received: May 03, 2016 Accepted: August 14, 2016 Published: September 15, 2016

Citation: Fauziah Abdul Rahman, Muhammad Ishak Desa, Antoni Wibowo and Norhaidah A. Haris, 2016. An improvement of Knowledge Discovery Database (KDD) framework for effective decision. J. Artif. Intel., 9: 72-77.

Corresponding Author: Fauziah Abdul Rahman, Malaysian Institute of Industrial Technology (MITEC), Universiti Kuala Lumpur, Persiaran Sinaran Ilmu, Bandar Seri Alam, Masai, 81750 Johor, Malaysia

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

According to Fayyad *et al.*[1], using Data Mining (DM) as a tool alone, failures in real environment since the analysis results not interpret as the whole picture of business perspectives although DM has been an established field[2]. It is also agreed by Wang *et al.*[3] that despite the maturity of DM, recent critiques state that DM does not contribute to business in a large scale. Usually the aims of DM are to develop a new approach or method. Datasets mined are abstract or refined from real problems or data, model and methods in DM systems usually predefined[4]. The key successful applications of DM are collaboration and knowledge sharing among frontline users and technology experts in the organization. Although there are countless researchers working on designing efficient data mining techniques, methods and algorithms but unfortunately, most DM researchers pay much attention in developing DM models and methods[5]. However, nowadays researchers with strong industrial engagement realized the need from DM to Knowledge Discovery Database (KDD) to deliver useful knowledge for the business decision making. In the real world scenarios, challenges always come from specific domain problems, hence the objectives and goals of applying KDD are basically problem solving to satisfy real user needs[6]. The KDD framework life cycle representation of DM process seems to become more dominant is a tool that enables one to intelligently analyze and explore extensive data for effective decision making. 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[7]. Using statistical method, sometimes one can find patterns are not significant in reality. Few researches were known to date about the usefulness of applying DM or applying existing KDD frameworks in transportation related research such as trucks or tanker maintenance actions. Now a days, an information system has been used extensively in many logistics companies to support their business processes including the fleet maintenance and tire management. It is estimated that the tire industry worldwide generates around one billion new tires each year. The continuous higher demanding, influence manufacturing and be one of the demanding and highly competitive industry and because of that any savings in raw material costs even 1% is a significant gain. The tire maintenance process is the most important component of tire management solutions. Over the last several years, new approaches and technologies have been developed for the commercial motor vehicle market to help improve tire maintenance practices, including automatic tire inflation systems and various types of tire monitoring systems. Tire Management System (TMS) for instance is generally a worldwide usage system that helps to manage overall tire business, start from tire supply management, tire selection, tire maintenance, monitor tire operational and also tire data analysis, so that many logistics company can achieve maximum tire lifetime with maximum casing life for retread, decrease breakdown time with best safety and finally reduce cost and more profit. Generally, the decision made by many logistics firms on when to perform tire preventive maintenance actions such as replacement of new tire, rotating tire positions and rethreading tire, is based on the experiences and intuitions of their maintenance staff. It is argued that the decision taken in this way may not be cost-effective. However, within the context in Malaysian Logistics Company, most logistics firms including a service logistic company (ASL) as a case study is manually tracks its trucks tire maintenance, with inspectors testing pressure and tread, they manually writing down the results and carrying out the maintenance when necessary according the details on paper although they are using TMS. Currently, ASL is a major player in land and agricultural plantation development especially in the palm oil and rubber industries and the respective support services. Substantial tire maintenance data are usually logged by such fleet maintenance systems but the data inside the system are not fully utilized as input for decision making. In order to fully utilized the data and be useful for maintenance decisions, the data needs to be properly and efficiently analyzed where DDDM-KDD is applied.

MATERIALS AND METHODS

Evolution of data mining framework: Cross-Industry Standard Process for Data Mining (CRISP-DM) is the first generation of KDD before DDDM-KDD. Is a data centered-heavily depending on data itself or data methodology or data-oriented base framework. It's provides a non-proprietary and freely available standard process for fitting data mining into general problem-solving strategy of business or research unit. For a real data mining problem, there need both of the background knowledge from users and data miners. In one hand, the user's background knowledge is important. This background knowledge can be incorporated with the induction algorithm and used for evaluating the mined results. In solving the problems that come from specific domains problem in a real world, next generation framework, Domain-Driven Data Mining (DDDM) has been developed specifically highlight the importance of data and domain intelligence[8]. Fundamentally, DDDM was including domain expert and domain knowledge. Domain knowledge consists of the involvement of domain knowledge and experts. But usually in DDDM existing work often stops a pattern recovery which is mainly based on technical significance and interestingness which including objectives and subjective technical measures. Interestingness basically refers to the pattern of result or rules at the end of KDD and is unexpected or desired to expert and being useful or meaningful. Therefore, it is important to have the involvement of domain knowledge in each phases of DDDM framework. However, different user may have different measures of interestingness pattern. Therefore, interestingness is strongly depends on the application domain, expert knowledge and experience. Therefore, actionable pattern has been added instead of just interesting pattern. In business, actionable pattern is more important than interesting pattern. This is because actionable is refers to the mine rules or pattern that suggest valid and profitable actions to the decision makers. This framework included two technical measures or metrics which are objective and subjective measures. The objectives measures are based on statistical strengths or properties of the discovered rules (data from database) and subjective measures are derived from the user's belief or expectations of their particular problem domain[8]. In order to encode the domain knowledge manual method, that semi-automatic and automatic methods have been used[9]. The automatic method requires some knowledge discovery tools such as ontology learning, knowledge acquisition based on ontology and semantic web[10]. However, these popular methods provide a conceptual or mapping representation of the application domain mainly elicited by analyzing the existing operational databases. Hence the interesting patterns and actionable patterns are based on the technical interesting pattern which refer to data and user's belief (domain knowledge) in particular domain.

Data analysis: The DDDM-KDD is one of the KDD frameworks often used for this purpose. In this study, we apply DDDM for formulating effective tire maintenance strategy within the context of a Malaysian's logistics company, A Service Logistic (ASL) as refer to the Fig. 1. Currently, the company practices tire maintenance policies which were formulated based on the experiences of its mechanics. Tire maintenance data from its TMS were not fully utilized as input to the current tire maintenance policies.

Stage 1: Business understanding related with goals of analysis: The problem understandings are given by ASL regarding the critical problem faced by them. Tire was identified to be among the most contribution 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 find the estimation kilometer of preventive breakdown based on journey per kilometers that causes to each detached reason. The knowledge obtained from the journey of each 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 tries to decrease of using new tire or certain expensive tire brand by getting the lower tire price. According to the analysis report done by ASL using statistical analysis in 2008, three root causes of higher tire maintenance cost, truck condition (alignment) which because of lack of maintenance, tire brand usage which are expensive and road surface condition which coarse and gravel road effect on the tire surface which mostly came from cargo trailer and it remain same until now.

Stage 2: Data understanding: This stage is investigates a variety of descriptive data characteristics for instance, count

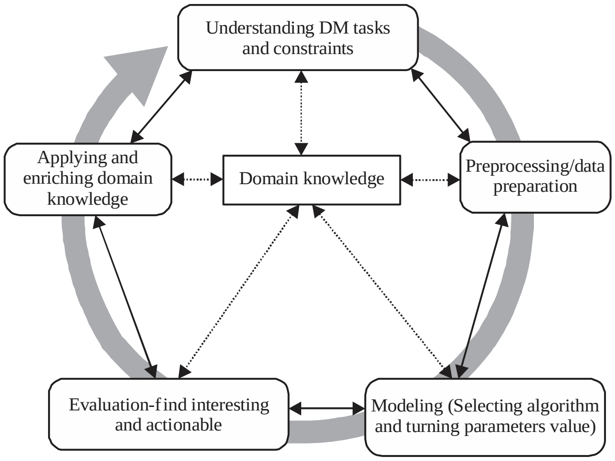

Understanding DM tasks and constraints
Applying and enriching domain knowledge
Domain knowledge
Preprocessing/data preparation
Evaluation-find interesting and actionable
Modeling (Selecting algorithm and turning parameters value)

Fig. 1: Domain driven data mining (DDDM-KDD) framework

of entities in table, frequency of attribute value and average values. The available data come from TMS system of ASL where all the tire transactions of three main depots have been recorded. Tire Management System (TMS) records all tire transactions including tire selection, depots tire maintenance, monitor tire operational and produce tire data report, to achieve maximum tire lifetime with maximum casing life for retread, decrease breakdown time with best safety and finally reduce cost and more profit. In this study the analysis was included 1,016 records from the time period of year 200 within tire transactions of three depots. Selected entities and records were based on objectives of analysis. Figure 2 shows that the significance level between attached trailer attached prime, journey, detached reason and tire brand where some of the correlation is significant and variables are linearly related or vice versa.

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. Selected entities and records consists of the name of trucks and trailers, detached odometer, attached odometer, detached reasons and attached tire positions. In addition, tire brands have been added into TMS attributes for analysis purpose was taken from Vehicle Maintenance System (VMS). This stage involves data cleaning activities. 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. The cooperation from domain knowledge and experts were needed for clear understanding. Data Cleaning (DC) has been done for modelling process.

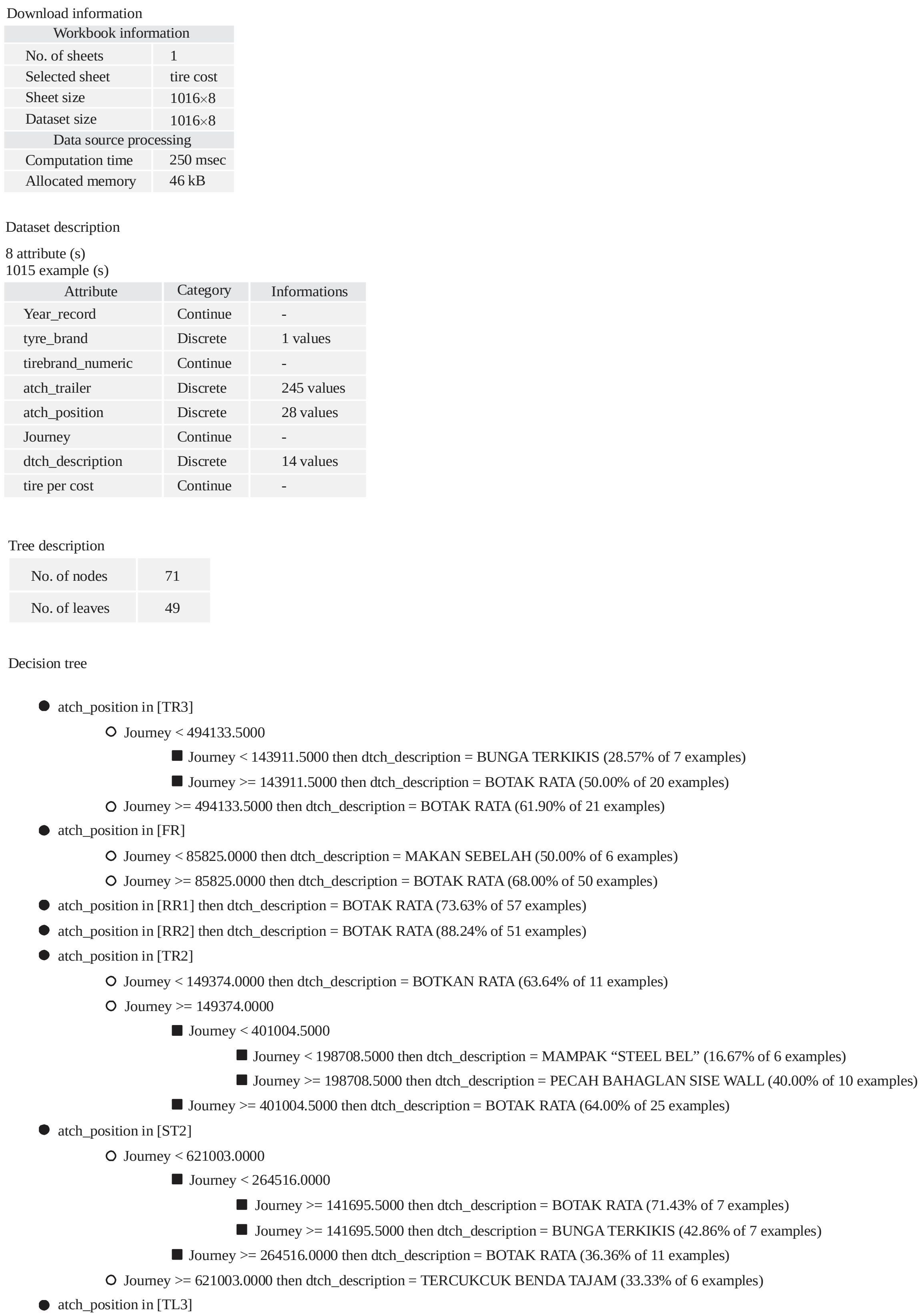
Stage 4: Modeling: Tanagra 1.4.40 open source data mining is a software tool to assess methods for DM problems. This method has chosen because of the significant correlation using Exploratory Data Analysis (EDA). Based on EDA, four attributes have been selected; tire brands, attached positions, detached reasons descriptions and journey. The data have been analyzing using a-priori association rule (AR) algorithm, unfortunately it did not produced any interesting rules. The same records have been analyzing using decision tree algorithm using C4.5 method and it produced few rules as refer to Fig. 2. The selection of DM methods was based on the literature review. The interesting rules were filtered by domain and experts in ASL company.

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 58,000 km 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 of journey >= 58,729 km until less than <132,692 km. Based on the domain expertise and domain knowledge, a new tire always can be use until 80,000 km 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. The 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 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 DR 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 or using the other DM methods.

RESULTS AND DISCUSSION

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 studies including an association rules and clustering techniques including optimization areas to have optimum decision making. In the real world scenarios, domain experts are slightly important for data validation in DDDM-KDD



Download information

Workbook information	
No. of sheets	1
Selected sheet	tire cost
Sheet size	1016×8
Dataset size	1016×8
Data source processing	
Computation time	250 msec
Allocated memory	46 kB

Dataset description

8 attribute (s)
1015 example (s)

Attribute	Category	Informations
Year_record	Continue	-
tyre_brand	Discrete	1 values
tirebrand_numeric	Continue	-
atch_trailer	Discrete	245 values
atch_position	Discrete	28 values
Journey	Continue	-
dtch_description	Discrete	14 values
tire per cost	Continue	-

Tree description

No. of nodes	71
No. of leaves	49

Decision tree

```
● atch_position in [TR3]
        ○ Journey < 494133.5000
                ■ Journey < 143911.5000 then dtch_description = BUNGA TERKIKIS (28.57% of 7 examples)
                ■ Journey >= 143911.5000 then dtch_description = BOTAK RATA (50.00% of 20 examples)
        ○ Journey >= 494133.5000 then dtch_description = BOTAK RATA (61.90% of 21 examples)
● atch_position in [FR]
        ○ Journey < 85825.0000 then dtch_description = MAKAN SEBELAH (50.00% of 6 examples)
        ○ Journey >= 85825.0000 then dtch_description = BOTAK RATA (68.00% of 50 examples)
● atch_position in [RR1] then dtch_description = BOTAK RATA (73.63% of 57 examples)
● atch_position in [RR2] then dtch_description = BOTAK RATA (88.24% of 51 examples)
● atch_position in [TR2]
        ○ Journey < 149374.0000 then dtch_description = BOTKAN RATA (63.64% of 11 examples)
        ○ Journey >= 149374.0000
                ■ Journey < 401004.5000
                        ■ Journey < 198708.5000 then dtch_description = MAMPAK “STEEL BEL” (16.67% of 6 examples)
                        ■ Journey >= 198708.5000 then dtch_description = PECAH BAHAGLAN SISE WALL (40.00% of 10 examples)
                ■ Journey >= 401004.5000 then dtch_description = BOTAK RATA (64.00% of 25 examples)
● atch_position in [ST2]
        ○ Journey < 621003.0000
                ■ Journey < 264516.0000
                        ■ Journey >= 141695.5000 then dtch_description = BOTAK RATA (71.43% of 7 examples)
                        ■ Journey >= 141695.5000 then dtch_description = BUNGA TERKIKIS (42.86% of 7 examples)
                ■ Journey >= 264516.0000 then dtch_description = BOTAK RATA (36.36% of 11 examples)
        ○ Journey >= 621003.0000 then dtch_description = TERCUKCUK BENDA TAJAM (33.33% of 6 examples)
● atch_position in [TL3]
```

Fig. 2: Decision tree method (C4.5) produce rules

methodology. However, researchers have difficulty experienced in term of long time doing the data preparation and modelling because waiting for the feedback for each phases from the domain expert that yet produced an inaccurate result. Therefore, an automated method of producing interestingness from domain experts should be use. Additionally, 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.

CONCLUSION

It is belief that the results can be improved by increasing the number of variables and records and an automated method for converting domain knowledge into DM process must be put into consideration. Using DDDM-KDD, it has been developed specifically highlight the importance of data and domain intelligence. Domain knowledge consists of the involvement of domain knowledge and experts. However, from the results, the patterns of rules are not informative and useful for ASL company to make an action based on the rules. An interesting rules is very important for them to support business decision making and operation. Therefore, the studies on DDDM have been extended to effective and practical methodologies for Actionable Knowledge Discovery (AKD). The research will aim to improve the Actionable Knowledge Discovery (AKD) framework which previously, most of the research focused on technical interestingness (technical data inside the system) rather than the business interestingness (business interesting patterns). It is important to get balance actionable knowledge based on technical and business interestingness in order to support effective or optimal solution for business action.

Furthermore, DC processes in data preparation also another area need to be focus instead of using DM methods. In future, we also need to explore other DM methods instead of classification and AR method, make comparisons so that the analysis goal can be achieved.

ACKNOWLEDGMENTS

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

REFERENCES

1. Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1996. From data mining to knowledge discovery in databases. AI Mag., 17: 37-54.
2. Huang, A., L. Zhang, Z. Zhu and Y. Shi, 2009. Data Mining Integrated with Domain Knowledge. In: Cutting-Edge Research Topics on Multiple Criteria Decision Making, Shi, Y., S. Wang, Y. Peng, J. Li and Y. Zeng (Eds.). Springer, New York, USA., ISBN: 9783642022982, pp: 184-187.
3. Wang, W., H. Chen and M.C. Bell, 2005. Vehicle breakdown duration modelling. J. Trans. Stat., 8: 75-84.
4. Zhu, Z., J. Gu, L. Zhang, W. Song and R. Gao, 2009. Research on domain-driven actionable knowledge discovery. Commun. Comput. Inform. Sci., 35: 176-183.
5. Cao, L., D. Luo and C. Zhang, 2007. Knowledge actionability: Satisfying technical and business interestingness. Int. J. Bus. Intell. Data Mining, 2: 496-514.
6. Luo, D., L. Cao, C. Luo, C. Zhang and W. Wang, 2008. Towards business interestingness in actionable knowledge discovery. Proceedings of the Conference on Applications of Data Mining in e-Business and Finance, June 26, 2008, Amsterdam, The Netherlands, pp: 677-692.
7. Barai, S.K., 2003. Data mining applications in transportation engineering. Transport, 18: 216-223.
8. Chen, C.Y., T.Y. Chou, C.Y. Mu, B.J. Lee, M. Chandramouli and H. Chao, 2005. Using data mining techniques on fleet management system. Proceedings of the 25th Annual Esri International User Conference, July 25-29, 2005, San Diego, California, pp: 1-10.
9. Gomez-Perez, A., 2004. Ontology Evaluation. In: Handbook on Ontologies, Staab, S. and R. Studer (Eds.). Springer, Berlin, Germany, pp: 251-273.
10. Cao, L. and C. Zhang, 2007. The evolution of KDD: Towards domain-driven data mining. Int. J. Pattern Recognit. Artif. Intell., 21: 677-692.