Predicting Passenger's Behavior by Using Association Rules in Data Mining

الاستفادة من تقنية استخراج المعلومات في معرفة سلوك و خصائص مستخدمي المواصلات

By

Wesam Mohammad Alnusairat
Student ID 120129

Dissertation submitted in partial fulfillment of
MSc Information Technology Management
Faculty of Engineering & Information Technology

Dissertation Supervisor

Dr. Sherief Abdullah

Jan-2016

DISSERTATION RELEASE FORM

Student Name	Student ID	Program	Date
Wesam Mohammad Alnusairat	120129	Information Technology Management	11/Jan/2016

Title

Predicting Passenger's Behavior by Using Association Rules in Data Mining.

I warrant that the content of this dissertation is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that one copy of my dissertation will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make that copy available in digital format if appropriate.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my dissertation for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

Signature

Abstract

Data mining is considered as one of the most powerful and contemporary tools in extracting massive and valuable data and information from several standard database formats that all different kinds of public and private organizations use in their daily operations and processes. In this thesis, the data type used is only one of millions of other possibilities and examples around the world where data is a real need for enhanced business operations and improved service offerings.

This dissertation offers a holistic proposal for utilizing data mining tools and technologies with the objective of collecting big volume of personal and demographic information about anonymous everyday public transport users. This will contribute well to the vision and strategy of one of Gulf Cites. Furthermore, it will assist the Road and Transport Authority in enhancing the levels of its service reliability and dependability while improving its security and safety measures and surveillance systems which offer priceless benefit to local community, its stability and prosperity.

Dealing with this huge volume of transactions could be very challenging. Chapter: Data Preparation in this thesis explains in a lot of details how data can be gathered all the way to the phases where data is organized, processed, reported and accessed. The data pre-processing phase touches on several topics such as handling the missing data, duplicating and writing some back-end procedures in order to modify the logic of current graphical data and move it to different table structure to be more fit for purpose.

As per the research conducted in this theses, it has proved to be very challenging to find any similar studies in the whole Middle East and Africa that look into and analyses the behaviors of public transport passengers using a data mining technology. This makes this thesis more unique and hopefully many other specialized studies and researches in the region will follow and contribute to what is fast becoming a compelling need in our today’s business management and operations in the whole world.

خلاصة

تعتبر عملية استخراج البيانات بأنها واحدة من أقوى الأدوات في كيفية استخراج المعلومات واسعة النطاق والمتكونة من عدة مصادر مختلفة في قواعد البيانات، وتمتاز بخاصية توحيدها لجميع الأنواع المختلفة من المؤسسات العامة والخاصة و تستخدم في العمليات اليومية المختلفة.

تقدم هذه الأطروحة اقتراح شامل لاستخدام أدوات وتقنيات استخراج البيانات بهدف جمع حجم كبير من المعلومات الشخصية والديموغرافية حول مستخدمي وسائل النقل العام اليومي الغير معرفين والذين لا تتوفر معلوماتهم في النظام. وسوف يسهم ذلك بشكل جيد لرؤية واستراتيجية هذه المدينة الخليجية. وعلاوة على ذلك، فإنه سيساعد مؤسسة المواصلات في تعزيز مستويات موثوقية الخدمة وزيادة الاعتمادية وتحسين تدابير الأمن والسلامة، وأنظمة المراقبة، وبالتالي فإنها تقدم فائدة لا تقدر بثمن للمجتمع المحلي والاستقرار والازدهار.

التعامل مع هذا الكم الهائل من المعاملات عادة ما يكون صعب للغاية. يشرح إعداد البيانات في هذه الأطروحة في الكثير من التفاصيل من عملية تحضير البيانات وتتبع كل وسيلة من وسائل معالجة البيانات حيث تم تنظيم البيانات ومعالجتها والإبلاغ عنها والوصول إليها. وقبل معالجة البيانات في عدة مواضيع، تشرح الأطروحة العديد من الحالات العملية في كيفية التعامل مع البيانات المفقودة، والمكررة وكتابة بعض الإجراءات للجهة الخلفية من النظام من أجل تعديل منطق البيانات الرسومية الحالية ونقلها إلى بنية مجدولة أخرى لتكون أكثر ملائمة لهذا الغرض.

وباستخدام بنامج ويكا المتخصص في أدوات استخراج البيانات، أودت التجربة على بيانات حقيقية إلى الكثير من النتائج المثيرة للاهتمام التي يمكن أن تكون مرتبطة مباشرة بسياسات وأنظمة جديدة. أيضا، وقد صيغت أكثر من خمسة وعشرون قانونا للسياسات فيما يتعلق بسلوك الركاب من شبكة النقل العام في هذه المدينة نتيجة لهذه الدراسة.

وفقا للأبحاث التي أجريت في هذا الأطروحات، فقد ثبت أن هناك صعوبة كبيرة لمحاولة العثور على أي دراسات مماثلة في الشرق الأوسط وأفريقيا حيث الاهتمام بتحليل سلوك الركاب في وسائل النقل العام باستخدام تكنولوجيا استخراج البيانات في هذه المنطقة. وهذا ما يجعل هذه الاطروحة فريدة من نوعها ونأمل أن تلحقها العديد من الدراسات والأبحاث المتخصصة الأخرى في المنطقة والتي سوف تساهم في سرعة التحول والاهتمام أكثر في تطبيق تقنيات استخراج المعلومات.

Acknowledgements

Firstly I would like to thank my late Father for His Grace, benevolence and for giving me the determination to overcome many trying moments to pursue my dreams.

I acknowledge with extreme gratitude the professional supervision from Dr. Sherief Abdullah. His attention to every detail and academic precision provided me the necessary direction and focus for my study.

I am also most grateful to my mother and my family for their belief in me and their constant prayers.

My heartfelt thanks are extended to my Wife, who took time off from her busy schedule to proofread my drafts.

To my wonderful children, thank you for bearing with me and my mood swings and being my greatest supporters.

Declarations

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Wesam Mohammad Alnusairat)

Contents

Chapter 1 Introduction

This chapter provides an overview about the whole structure of this thesis. It starts with providing some background about Data Mining in general and then explains the main aim and objectives of this thesis. The fundamental quest of this thesis will be also presented as part of this chapter together with a summary introduction that explains the overall structure of this thesis, its chapters and topics.

Aims and Objectives

The Gulf City is considered as one of the fastest cities in the world in terms of the pace of development. It deservingly ranked within the top five fastest growing economies; thanks to its ever growing trade and tourism volume. The Government has decided to create a dedicated authority to manage the transportation challenges, with the mandate of enhancing and maintaining road infrastructures and developing well connected and integrated public transportation systems. This was seen vital for the attainment of this Government's vision which cannot come true without having carefully established plans and, supporting policies and legislations and smart technologies. With this in mind, the transportation organization has adopted new technologies and innovative approaches and leveraged global best practices and standards in road and transport areas. This thesis sheds light on one of the most evolving and modernized ideas that this organization has embraced in its effort towards innovation and service integration by using a smart cards. These smart cards are a unified smart cards that allows residents and tourists to use in all public transportation services including Buses, Metro, Waterbuses, Taxis, Tram as well as paid parking. It is supported by a well-designed and integrated 'touch & go' systems that simply reads a pre-paid card on any of the entry points of the above services and then deducts the fare amount in a precise and reliable manner. It is quick and easy to use and enables passengers across the gulf city to use several public transportation mediums without having to have cash on their wallets.

As an organization, the transportation organization consists of several functions and segments, such as road authority, public transportation, rail, parking and marine. With more than 800,000 daily users of the smart cards, it is, without a doubt, important to study and analyze how these users behaved in the past in order to better predict their future actions and behaviors.

However, the majority of the public transportation passengers in Gulf City are anonymous users. The reasons for that are not limited to:

Commercial reason: As per the rule of government, the passenger should pay 30 AED fee as a registration fee, because the registered users enjoy a lot of features such as refunding the remaining amount, blacklisting the lost cards and many offers and discounts. This 30 AED fee prevents a lot passengers from registering his/her personal information of the transportation card.

Privacy reason: this is related to the culture and mentality of the people that not all of passengers are willing to share their personal information although the registration is free of charge.

Technical challenges: this is very tricky factor because getting personal information required data integration with the emirates ID card, and this integration requires a lot of system tuning to stabilize it and improve it.

Indeed, they are many reasons to minimize the passengers from registering their personal information. This dissertation, will help the transportation organization to gain more demographic information of all anonymous passengers, and this will improve the decision making process and enhance the strategic planning activities. This is the main problem this thesis is trying to solve is to predict the demographics of unregistered users using their behavior and how data mining can help by learning from the small subset of users where demographics are provided.

Let assume that the transportation organization wants to inform public transportation users (e.g. about a delay in metro or out-of-service bus). It can be seen costly and disruptive if this message goes out to all users. This would give simple use case about the importance of having a personal information of the users, so that the message could be delivered to only the affected users.

Another example where such an analysis can be crucial is planning expansions. Suppose the transportation organization is planning to open a nursery in one of the metro stations, how the transportation organization can determine which metro station is the best for such matter? This thesis basically studies the behavior of the personalized cards information, in order to understand the unknown user's behavior. Data mining techniques will be used to build predictive models and patterns that can help in predicting the demographics from metro user behavior.

Overview about Data Mining

Across the world, there is a tremendous number of IT systems and software solutions that exist to serve our human needs and help us in doing things quicker, easier and better, whether through direct or indirect real interaction with the human being. As a result of this interaction and communication between information technology systems and humans, tons of billions of light-speed transactions were established and stored on hard disks and databases. In most cases, users pay less attention to how the massive amount of data generated by these systems can be benefited in the areas as long as the business or the operations are working well in line what there were created for. Nevertheless, recently this idea has been changing in some parts of the world supported with some major forces such as market competition, economic development, and technology and communication advancement and consumer demands. Many decision makers and CEO's have started to realize that their business survival is never as easy as it used to be in the old days as competition is getting bigger and harder. As a way of survival, they found out that one of the best ways to maintain their competitive advantage and even grow it is through opening their lockers and start looking into their old data that they managed to accumulate over the years.

More than any time ever before, data mining and data discovery are increasingly becoming the new science that uses advanced tools, algorithms and techniques with the aim of mining big data and providing very valuable information that may help understand past events and patterns and anticipate future trends and behaviors. Furthermore, Data Mining can be utilized and leveraged fin most of business areas and domains, including marketing, technology, sales, health, operations and even more.

Data mining can be easily integrated with other sciences and fields and significantly enhance the quality of services we receive every day in our life especially when it is well supported by effective systems and procedures such as machine learning, data discovery, big data and distributed database. All these connect with each other to make effective use of the huge volume of the data that would otherwise sit on companies' database servers completely ignored and unutilized.

The interesting thing of studying data mining is its direct relation with one of the most contemporary topics in business is Big Data. Because big data as a science and practice is

expanding incredibly with all business fields and aspects such as biology, ,health care, pharmaceutical, aircraft, engineering, and many other complex domains. As such, [Wu 2014] has provided a new platform called "Big data Mining Platform" with three-tier structure in order to handle and deal with the challenges that are faced when mining big data.

Research Questions

The main goal of the thesis is to answer the following research questions:

Using the subset of registered users, can we use data mining techniques to build a predictive model about user demographics? In other words, can we predict the demographics of unregistered users using only their behavior of riding the metro?

Are there patterns in metro riding behaviors that distinguish different demographics such as gender and nationality?

Structure of the Thesis

The remaining of the dissertation is arranged as follows:

Chapter 2 will include the literature review about the data mining and its usage on the public transportation domain. Another related topic will be discussed on this chapter such as smart card, and the tool used on the experiments. Chapter 3 will explain and analyzes the collected sample and the dataset used in the experiments, and how the data prepared and preprocessed on details. Chapter 4 will cover the experiment results and the conclusion of the thesis and the future work will be described in the last chapter.

Chapter 2 Literature Review

This chapter provides a highlight about data mining, big data, Apriori algorithm, and mining transportation data in other implementations in the world. Also, it describes the used data mining tool in the research.

Data Management

Most of the current researches and studies clarified and proofed the positive impact of paying attention to the data. Data is the output of any running business and projects. Without storing the data for a long period, many organizations might operate well, but for how long?

Abdul Rahman et al. (2011) clarified how important is managing the logistics in real life, and with stressing and focusing on the science of data discovery and data mining, management of the logistics would be better, they proofed that logistics and managing transportation can be done with normal computer systems, but after sometimes, the failure could happen especially with the complex projects, but with building a knowledge discovery system based on data mining and knowledge discovery, this can avoid a lot of mistakes that the normal legacy system can catch.

As we are talking about the science of data management, we need to spotlight the big data world and its contribution on the enhancement of the logistics and transportation.

The concept of the big data is to focus of any piece of information regardless the value nor the size of it. Any bit of information is very and very important." In order to adapt to the multi-source, massive, dynamic Big Data, researchers have expanded existing data mining methods in many ways, including the efficiency improvement of single-source knowledge discovery methods (Chang et al., 2009).

Fiozia et al. (2013) explained that in our world of Information technology, we can get a great benefit of the sensors and smart agents that installed on the cars, sensors on the road, multi-agent devices, constant mobile networks with 3G or 4G, and virtual environment or clouding system to gather the information and provide an artificial and analytical tools on the clouding storage with high performance resources. Hence, small vehicle, big vehicle, road signals, road sensors,

smartphones are big nominee for feeding the big data environments. The only challenge here is the proper powerful platforms and servers to perform well with this kind of transaction volume, and this one of the challenges that faced on this research which is the analyzing huge volume of the transactions.

From above challenge, the idea of the business intelligence is taking place, and start being considered from most of the business decision across organizations. Social media analytics, social network analytics like Facebook, twitter, mobile virtualization, prediction and data mining, dash boarding and more will be the normal result of applying a strong business intelligence system in the organizations (Chen et al., 2013).

Zhong1, et al. (2010) managed to build a special data warehouse environment in order to analyze the behavior of a cargo business in all transportation modes in China. The data warehousing was based on the mulita dimensional database contains of data transfer using ETL and creating the multi-dimensional data set, and the result was a time flow analysis for all shipment and containers. Shaw et al. (2003) got benefit from using GIS technology to retrieve important information for the land and analysis purpose, and they build a prototype to enhance the relation between the land and the transportation by analyzing the GIS information.

Smart Cards

I am going to highlight the usage of smart cards on the real life nowadays, because this the main source of data in this study. We used the transactions generated by the smart card on two transportation modes: Bus and Metro. However, the idea of unifying the transportation payment is not very new, Ronald F. Cunningham (1993) made a research in New Jersey city in new York to propose a system that allow people to pay the parking fees through a smart card.

However, the idea was not very successful as the culture of the payment was not very advanced and accepted by the society at that time. Kusakabe et al. (2014) made a very similar study to this research by analyzing the data coming from the smart card and understand the behavior of the trips, and not the passenger. However, the model build on their proposal was based on passenger survey to validate the results and compare with a survey. The good thing on their analyzes was the duration of the study as it was built on a 20 months of smart cards transactional data, while this study analyzed only one month of data, which give it less accuracy. Also, they mentioned

that the average number of trips per day was around 3000 trips while in our study reaches to 34,000 trips per hour. On the other hand, Dang et al. (2014) made a research to study the different type of cards used in public transportation, which is energy circle cards.

The main purpose of using this card is because it is very friendly to the environment, and keeping most of the advantages and specification of normal smart cards. This study was mainly supporting and encouraging using these type of cards as it has features of sending the urban transaction to the transportation department, and this department can easily manage the logistics and transportation from the stored information by these cards. However, the research doesn't talk about the compatibility and the failure chip ratio of these cards and wither it can handle all security keys to be stored on it. I think we also need to study the efficiency of these cards and the stability on such strong climate such as the gulf area climate in the summer season, which the temperature reaches up to 50 degrees most of the time.

In order to try to solve or mitigate one of the most critical problem that faces the biggest cities in the world, we can utilize the technology of smart cards and the science of data mining, to develop a model that incredibly helping in reducing the traffic jam in Beijing in China. The algorithms used on this research are clustering algorithms that used: three-hidden-layers Neural Network, K-Nearest Neighbor, C4.5 Decision Tree, K-Means. After using all these clustering algorithms, they managed to classify the transportation pattern regulations with a higher accuracy level using the k-mean clustering algorithm, and they managed to cluster the travel behavior (Ma et al., 2013).

Pelletier et al.(2010) clarified the usage of automatic fare collection system and its important role of the financial aspect, and its usage in many places, and it clarifies some of the disadvantages of using this system, such as the information access, and the cost wise of development and change request of these systems, due to the complexity of the software development. In our system in City, we haven't faced these problem even though we face another problem that the public authority couldn't find any trusted company or vendor on the Middle East regain, and speakers Arabic to facilitate the communication and understand the terms and conditions. And this causes a lot of problem due to conflict and misunderstanding on the contracts and maintenance responsibilities. I want to finish this section with a small advice of the smart card: "smart cards don't leave the house with it" (Husemann 1999).

Data Mining

Yuan et al. (2014) provides a real example of the importance of using the data mining of offering the accurate factors about the traffic jam on the normal and emergency situation. The nice thing of their system is that they can simulate a lot of different system in one system, then they can study the impact of mining the data with the result that would help on predicting the traffic. Another important thing is the way that they are capturing the traffic information, which is via video devices that reads input from online and real data to have a real idea about the input of the system. This will reflect the output of the prediction information.

Wang et al. (2014) build their case study with a similar condition to gulf city transportation system in terms of volume of the data set and the volume of the daily transactions. They also were very concerned about the customer satisfaction, and they used the CRM as a channel method to hear their customer feedback about how they are satisfied or no, because they cannot distinguish the customers by their behavior or their characteristics. Distinguish the customer using CRM is not an efficient at all due to the huge number of customers whose are not approaching the CRM and record their behavior.

However, they have used the data mining clustering to split the customer into groups related to their movement: short movement, medium movement, and large movement. The algorithm used on their analysis is FCM (Fuzzy c-means) clustering. Park (2014) clarified that the science of data mining is not just a theory anymore, in the one of the biggest and most dynamic city in the world, New York, the transportation department created and developed a full system called DMS, data mining system to be an essential part of their daily life operations, while a lot of transportation authorities in many cities they have not heard about the data mining and its importance yet. They used the data mining to analyze a very challenging data such as the data came from CCTV as well as the normal transitional data that come from the cars and signal sensors. Also, for the data mining world, we can see a lot of fuzzy tracker algorithms used for simulation in order to provide a solution for transportation network on the labs (Gel 2006).

(Haluzova 2008) shows another example of using one of the most famous data mining which is the association rule in order to come up with rules that will help them on their analysis. The same as this study and in order to use the association rule, a lot of data preparation work is required, for example: they have changed the values of the attributes for the days in the week from 1 to 7,

gave a ranking for the time in the day from morning till the evening time. Cheng et al. (2010) explained an another example of data preprocessing to prove that using the same data collected by passengers and vehicles are irrelevant for data mining processing as is. It requires a lot of reprocessing techniques that includes: data cleansing from missing data, corrupted data or duplicated data, then this requires data integration and data migration into intermediate level or staging environment where the mining of the data takes place. Wuhan et al. (2008) found a very interesting rules for a transportation data after applying association rule technique on it. The three rules found are:

R1:L5=> V1 10%, 16% indicates that the possibility of running a red light in place L5 is 16%, and support degree is 10%.

R2:E2=> V2 31%, 42% indicates that the possibility of the driver that education is senior middle school running a red light is 42%, and support degree is 31%.

R3:T1=> V1 36%, 40% indicates that the possibility of running a red light in the morning is 40%, and support degree is 36%. ”

The other interesting thing of the usage of data mining that some authorities are start thinking about mining the data that come from the bikes and not all from the vehicles, because the usage of bikes on the cities are getting increased(Vogel 2011). Another benefit of using the data mining in the transportation work is to improve the safety and security for the roads and passengers.

Weka

The usage of data mining techniques is getting increased. It is being adopted by most of the critical aspects in our life that requires data analysis. However, we can expect to use data mining in many fields, but using the techniques of data mining in healthcare domain is very challenging, especially if it is about using the techniques to predict the breast cancer reoccurrence. They simplify this object by focusing on the main concept of data mining which is analyzing data from a data set in order to predict and expect some result related to the medical tests.

Jethi et.al (2015) analyzed the data set of the breast cancer, and they have used two data mining techniques: clustering and association rule, in order to reach to more accurate figures. However, increasing the number of techniques to increase the accuracy of result sounds promising, but it requires more research to proof the results to cover more circumstances. Eight years ago,

Othman (2007) tried to use Weka tool to use many classification methods for breast cancer data. They used decision tree algorithms and they found that 151 out of 175 ladies got classified correctly with accuracy level around 85%. Frank et al. (2004) used Weka tool to analyze bioinformatics data.

However, they mentioned one of the biggest weakness of Weka tool which is the performance issues, and they have proposed using a distributed environments and to split the data set among them in order to mitigate this bottleneck, and in order to increase the performance of Weka execution time required for performing such analysis and provide the results. Stevanovic el. Al (2012) explained another example of using Weka by using classification algorithm, in order to analyze the websites user sessions traffic and session logs, in order to find out the classification of websites user experience. They start their data preprocessing by collecting all the website logs from the web server and keep it on access database, then they have labeled the session into classes and divide them into the training data set and testing data set. However, they have reached to classification accuracy near 100%, this means the classification result is as expected, and there is no important contribution with this result.

I think it is a good idea to use the data mining to analyze the websites logs, but this requires more attention on the preprocessing phase because the logs are considered not to be a structured data, and it is very critical to work with it.

Sharma et. Al (2012) used the Weka tool for K-means clustering algorithm in order to analyze crop yields records. They have done their experiment only to proof that Weka is capable of clustering the data set by setting the proper K-means parameters and reach to the most proper n value for a number of clusters. However, they have mentioned that as a weakness of K-means by the requirement of setting the value of N before start the analysis, but this is not a weakness from my point of view due to the logic of the K-means algorithm and the practical side of experiments and keeping trying different variables till reach to the closest result. one point related to the performance and time consuming when analyzing the heterogeneous data set, we can't depend on one data mining tools for all kind of data analysis, some tools are very strong while other tools are very weak in terms of the nature of the data set and verse vice (Angela et al., 2014).

Finley, Hall et al. (2009) stated that more than 64 critical projects used Weka for their analysis as per the Weka site, and they mentioned that Weka is the best open source data mining tool.

Chapter 3 Methodology

The used methodology and the process of the data collection will be explained in this chapter, and all the levels of data preprocessing, because the data was manipulated many times from the time is selected from the legacy system till the end with the final data set that used in the experiment.

Research Approach and Techniques used

In data mining, association rule learning is a general and well-studied technique for determining related relations between factors and variables in big databases, and this could be applied on many example in our daily life. The most famous example of using association rule is the usage of market baskets. For example: the rule {Tomato Catchup, Potato}=>{Oil} found in many data set of any supermarket. This would show if a customer buys Tomato Catsup and potatoes on the same basket, he or she most probably will by Oil on the same basket. This kind of data can be analyzed and help to create as the basis for decisions makers to help them making their marketing decision. Sales and product arrangement are good example of getting benefit of these technique. From the heart of association rule technique, I found that our transportation data set is very similar to the above example of market basket example. The transportation data is very good example of using association rule, because it behave the same function and behavior of this technique. For example: if customer travel from point A to point B, then most probably his gender is Male. Hence, and because of the fact that the Apriori algorithm is a classic algorithm for learning association rules, it is been decided to use this algorithm in analyzing the gathered data set as it will be illustrated on the next section. Apriori algorithm is designed to work on large data sets containing transactions (for example, groups of items or details of a website hits).

Apriori works an iterative method known as level-wise search, where k-item sets are used to discover (k+1)-item sets. The first set of frequent 1-itemsets is found by look over the database to collect the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted as L1. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-item sets can be found. The finding of each Lk requires one full scan of the database.

At final iteration you will end up with many k-item sets which is basically called association rules. To select interesting rules from the set of all possible rules various constraint measures such as support and confidence is applied.

Apriori Pseudocode

Apriori (T, ε)

$L_1 \leftarrow \{$ *large 1-itemsets that appear in more than* ε *transactions* $\}$

$k \leftarrow 2$

while $L_{k-1} \neq \varnothing$

$C_k \leftarrow$ *Generate* (L_{k-1})

for transactions $t \in T$

$C_t \leftarrow \text{Subset}(C_k, t)$

for candidates $c \in C_t$

$\text{count}[c] \leftarrow \text{count}[c] + 1$

$L_k \leftarrow \{c \in C_k | \ \text{count}[c] \geq \varepsilon\}$

$k \leftarrow k + 1$

return $\bigcup_k L_k$

Dataset Description

The initial data set is represented a full month of real records of the smart card transactions, with all kind of transportation modes includes: Bus, Marine, Metro and Parking transaction, which means all kind of transportation happened using the smart card in the month of February 2015 was captured in this sample.

Indeed, it is hard to assume that this sample includes 100 percent of all transaction, because still there is a possibility of losing some transactions from the source to the destination, due to the operation challenges and the technical faults. However, I made some statistics about the percentage of the lost data for the three previous months, and I found that around 98 percent of the actual transactions in all transportation modes that delivered to the clearing housing system. This percent is calculated based on a general comparison between the number of received data to clearing house system and the number received in the operation system.

The operation system was always having a bigger number, which leads to some transactions that lost, and this needs further study to eliminate this problem and to fix it. In addition to that, the original data was including all kind of transportation usage transaction type, such as bus CKI and CKO, metro CKI and CKO, marine CKI and CKO, and other kind of operational transactions such as card refund transaction, card replacement transaction, and reload and recharge, card sale transaction.

The data was collected of the full month of Feb 2015 for 28 days. The initial size of the data set was around 52 million records. The initial sample was very complicated, because it was including the E-purse transactions and the production transaction. The E-Purse of those cards having an amount of these cards and each time the E-Purse inside the card will be deducted from E-Purse depends on the fare should be deducted. While the product transactions comes with zero value, because when you buy the product one time and you use it as much as you want depends on the period of production days starting from one, seven, thirty days, quarter and yearly pass.

Extracting a 52 million of transactions was a very big challenge due to the huge size of storing and handling the data. 52 million transactions was around 19 Gigabyte of data and took around a week to gather the data together. The data set is collected from Oracle database version 11.2 enterprise edition and includes many data types like integers, varchar, number, Boolean and blobs.

Of course, not all the data is required on the analysis, but on the initial stages, it was very difficult to confirm which data is required and which is not required, that is the reason I have asked for all row data from the legacy system, and the plan was to handle and process the data later on far away from the operational database and depend on the progress of the analysis. That

is the reason why the data preparation took long time and effort from the total hours of this research. All data preparation phase will be explained in details on this chapter.

Data Preparation

The first challenge I have faced is how to handle the fifty-two million transactions as it was only in a one oracle dump file. A temporary Oracle database had been created to try to load the data. Oracle express version 11.2.0.1 had been installed in a windows machine 64-bit with 8 Gigabyte random access memory capacity.

However, after a several attempts of loading the data, the import job failed. The failure of loading the data was due to the huge size of the dump file. The other problem that the file wasn't easy even to open manually to try to split it or divide it. Then, I have returned back to the operational database to try getting less period of days.

The second round of extracting the data was based on four weeks' time. This time, the exporting of the data source was based on CSV format instead of Oracle dump format. This offers more flexibility of handling the data, then processing this kind of huge data. However, I have managed to import the data into different database base engine. I have done many searches about the best database engine of processing a huge data, and I found that SAP Sybase IQ database is one of the best powerful database technology.

Hence, I have decided to install Sybase IQ 15.4 version on windows 7 machine 64-bit. The result of importing the four weeks was very promising as I managed to load the data of a full month transactions of Feb 2015. That was about 52 million records on a database. I have decided this as a first step because of the strength and advantages of a database in terms of data perpetration and preprocessing, especially with working with this kind of volume of data. The volume of one-month data was decided to increase the accuracy of the result and analysis, because a bigger sample duration a more accurate data.

However, more than a month for this data set was very challenging and difficult due to the data set size required for this research. After finalizing the import activity, the performance tuning phase has started in order to increase the data retrieval time and increase the overall performance of the data.

Indexes, views, and stored procedure were created on top of the temporary tables to facilitate working with data. By this moment, the first phase of data preparation was finished and the data was ready to move to the next phase, which was data cleansing and prepossessing. Indeed, the next phase is required to prepare the data to be ready for using by the data mining tool. As there is no tool in the world can handle this number of records. The second reason of this phase was a more analytical point of view. On this stage, I had to finalize the visualization and confirming the final parameter that will be inserted to Weka for data mining process.

The second challenge I have faced is the presentation of data into the data mining tool. This problem appeared depends on the type of records placed on the legacy database. The original data was presenting on the concept of legs and trips. Each trip has a minimum of two legs: CKI and CKO. Hence, we will find two database records for each trip. For the benefit of better analysis, I decided to re-format the presentation of the trips to under one record. Hence, the output was the CKI information and CKO information on the same record. After this step, I have reached to the third challenge which is deciding the personal demographic that we need to build this research on them.

To answer this concern, we need to take the business point of view on to the consideration. Hence, after holding a meeting with the business people on the organization, I found that the most added value in terms of personal information to know for the passengers would be:

- Age
- Gender
- Marital Status
- Nationality

Indeed, from the business side, any personal and demographic information will definitely add a lot of values. Selecting the top four demographic criterial was made after consulting the strategic

planning department, because they are very interested to know more about these personal information, in order to help them in their future studies and the plans.

The process of the filter the data to include only those columns on the sample data.

The final sample data would be as below:

Card1,27,'MALE','SINGLE','India','SILVER','Station A'

Card2,15,'FEMALE','SINGLE','India','SILVER','Station B','Station C'

Card3,29,'MALE','SINGLE','Philippines','SILVER','Station D','Station E','Station A','Station D'.

By reaching to this level, the sample data is ready to start the analysis part which would take place on the next chapter.

Chapter 4 Experimental Analysis

This chapter aims to discuss the results of the data mining exercise executed on the sample data. This chapter will include four sections depends on the demographical information.

I have started the analysis with setting the minimum support count by 2, but it didn't retrieve any logical result, because the frequency of the item sets was very high. Then I kept trying to increase the minimum support count by 1 till I reached to the most proper value of the support that used in this analysis, which is 6. The reason behind this value is related to the huge number of transactions and the possibility of each iterations. In regards of the confidence value, and according to business decisions from the transportation organization strategic department, we all agreed to set the value of the confidence as per below formula:

Gender Confidence Value = Random Value + 10

Age Confidence Value = Random Value + 5

Marital Status Confidence Value = Random Value + 1

Nationality Status Confidence Value = Random Value + 1

This had been agreed after a discussion with the concerned department, and after sharing with them the random figures for each demographic category. For example: the random ration of some nationality is very low. For example: the X nationality passengers represents .5 % of the total passengers. Thus, we agreed that if we find a rule that shows the probability of using the metro by X nationality is more than 1.5% is very good achievements. Below are some rules as a result of the analysis that meets the support and confidence count per each demographic category.

Gender

After finalizing the sample data which was for the month of February 2015, I couldn't run the tool for the whole month, because I will not get the daily behavior of the passengers, which is the main objective of this research. Hence, I had to choose only one day to study the behavior of the passengers. But, sometimes the passenger behavior might change extremely depends on many factors. The first factor is the status of the passenger, resident, tourist or visitor. The other factor could be the special occasions like exhibitions and seminars, Eid timings, and Ramadan timing.

Another factor could be the multi-usage of one card, this might happen a lot of times, because many passengers can share the same travel card. Another factor could be the vacations of the schools and universities. For these reasons, I picked the month of February to eliminate a lot of above factors. However, the possibility of outliers and exceptions still be there. Hence, I have decided to use only the frequent users.

Then, I have defined the frequent users by those users they use the public transportation five times a week. Because their behavior is what I am looking for, in order to achieve the research goal. Then, I made some queries to eliminate the non-frequent users from the sample data. I almost removed the weekend's days from the sample data. On this case, and after I run some random queries, I verified that the behavior of the card holders is not affected by changing the date on the month of February.

Now, I will start with analyzing the behavior of the cards trips. The number of personalized passengers on the sample day was 27,221 card holders. And randomly, the percentage of male gender on this day is 68% with 20555 card holders, while the percentage of female passengers is 32% with 6666 card holders. The target was to use the data mining algorithms trying to increase the random percentage of the passenger gender. The result was as below:

If the CKI from the second leg of the trip from Metro Station A, then, the possibility of passenger gender is 94% MALE. If the CKI transaction of the trip was from bus stop B, then, the possibility of the gender of the passengers is 94% MALE. Also, 87% from passengers those starts their journey from Stadium metro stations are MALE.

I found another case that if the trip has two legs, the first CKO happened in metro station C and the second CKI from the second leg happened in the same metro stations, on this case, the possibility of the passengers to be a MALE is 85%. From the previous figure, we found that most of the employees those working on Area X and they are using the public transportation are MALE.

On the other side, I found that the percentage of 69% to be a FEMALE, if the passenger checked out from metro station C as their third leg. Also, if the transfer trip happened on metro station D, the possibility of the FEMALE passengers are 56%. As we can see from the sample, the random

probability of passengers is 78% as male, but with using the data mining software, I have succeeded to raise this ratio from 78% to 94% for such metro station. Also for the female gender, I have found that for such metro stations, the percentage of female passengers is very high.

Now, I will just give one example of the best utilization of above findings. Suppose that one investor needs to open a Ladies and Gents salon on one of the metro stations, or some shops selling female or male staffs, above findings will be very sufficient in terms of decision-makers, and choosing the most proper location instead of relying on the random figures only.

Age

Getting more accurate values related to age would very interesting for decision makers and strategic planners. On my study, it was very difficult to take each passenger separately. I tried executing the data mining tool, but I couldn't figure out any important data from my data set.

Then, I have decided to reprocess the sample data and change the data related to age into interval attributes. The values of the interval have been decided according to business needs and requirements. Thus, all attributes are divided into 3 intervals as below:

Young

Senior

Child

Even with creating those three intervals, the analysis was very challenging because of the ratio of each interval. As per the business, the age of passengers to be considered as young was from 18 years to 45 years. And the senior people ages above 45 years and the child are less than 18 years.

By these intervals, and if we calculate the percentage of each interval randomly, the young passengers represents around 88 percent of total passengers, and the senior represents 7 percent and the child passengers were only 5 percent out of the total public transportation passengers in the gulf city. From these value, the challenge was very high, especially with both senior and child intervals due to their low occupation. However, after running the data mining tool, below shows the result of the findings:

If the passenger starts his second journey from metro station A, the percentage of the age group is 98% will as young. The Same result for the metro station E, the percentage is around 97% to be a young. If the passenger starts his trip from metro station F, so this means that the probability of passengers to forms a young group sits 97%. Also, if the passengers do a check-out from metro station G, the percentage of young passengers is 92%. On the other hand, the only major findings related to other age groups is the following: if the passenger does a check-out from Metro station T, then the percentage of child passenger is 43%, this is a major finding due to the low random number of child category passengers in the current public transportation scheme.

I was curious about this result, that's why I have visited that metro station, and I found that this area is one of the top areas in terms of high capacity of school distribution in this gulf city. This explains the result and give it more reliability. The percentage of child users are very high.

Hence, the authority would pay more attention in terms of security and safety on this metro station, and some restaurant/grocery for children they can start their business in this metro station.

In regards of the senior category, and after executing a lot of data mining iteration, I couldn't find any major findings or observation. This is because the low number of senior people those using the public transportation, or they are not focused on some area and they are all distributed equally on the public transportation stations.

Marital Status

The third demographic factor that this study trying to figure out is the marital status. The marital status is very important as per the input from the business due to its strategic importance and taking the proper business decision. Based on the figures from our sample, the marital status statistics is as follow:

The total number of personalized passenger on the sample day is 16,056 passengers distributes as:

SINGLE 98%

MARRIED 2%

DIVORCED 0%

WIDOW 0%

From above figures related to personalized passengers, we identify the complexity and the challenge of extracting any useful information for anonymous people due to the huge percentage of the single passengers.

I have discussed this result with the concern department and I found that marital status field on the personal application form is not mandatory, and the system would depend and use Single value as a default value for the non-filled application. Hence, one of the suggestion needs to be analyzed is the mechanism of changing the application form to force the people inserting more real information during filling the application of getting the personalized transportation cards.

Going back to data mining practice to figure out the marital status. Unfortunately, all of the rules related to marital status factor was a single, and I failed to find even a single rule talking about any other marital status group. However, I found some rules where the percentage of a Single group is higher than the random percentage which is 98%, but this represents less significant value, because there is no much difference between the random parentage and full percentage. Which make the result less importance in terms of business decision. I am listing all the rules whereas the percentage is greater than 98%:

If the trip checkout from Station C

If the trip check in from metro station F

If the trip check in from metro station C

If the trip check out from metro station D

If the trip check out from metro station N

However, most of the finding very close to the general statistics. I couldn't find any valuable rules for marital status because most of the users are single (98%). However, I found a lot of percentage of such metro stations are 99% single. But this can lead to thinking that the authority should pay more attention to stanch the families and married people and convenes them to use public transportation.

Nationality

The last demographic factor on this research is the nationality. This factor is the most one in terms of business needs due to its importance for decision making and strategic planning. I have used the same sample which contains 16,056 records. Below is the top 30 nationality with their percentage:

A	55.88
B	20.86
C	7.01
D	2.81
E	2.6
F	1.87
G	1.22
H	1.08
I	0.79
J	0.52

K	0.49	
L		0.32
M		0.25

From the random figures above, it is shown that the majority for the passengers are from nationality A, B and C with a different ratio. However, the target is to find a rule with a higher percentage than the random figure. After running the tool, I found out many interesting rules based on the nationality as below:

If the check out in the second leg happened on metro station D, the percentage of the nationality will be an A is 75 percent, while the random figure is 55 per cent. If the check out in the second leg happened on metro station C, then, the percentage of the nationality to be an A is 80 percent. The checkout happened on metro station F, then, the percentage of the nationality to be a B is 77 percent while the random figure for the passenger to be a B is only 20 percent. Even if the CKI in happened on metro station O, then 77 percent for the passenger is B passengers. If the checkout happened on metro station T, then the percentage of B nationality is 70 percent. If the passenger start his journey from metro station E, then the percentage of the nationality of the passengers is 60 percent is A.

Another important rule related to another nationality is the B nationality. If the passenger done a check out on metro station C, then, the parentage of the passenger to be a C is 16 percent while the random figure for C nationality is only 7 percent. Even for D nationality, using the data mining tool we can know that if the passenger CKO in metro station L, then the percentage of the passenger to be D is 14 percent, which is five times the random figure.

Even though the A passengers represent the majority of nationalities of public users, this study found that some metro stations passengers bigger than the random ratio, this enhances the likelihood of approaching the nationalities for any future business and events.

Chapter 5 Conclusion and Future Work

This section highlights the overall conclusion of this thesis providing clear answers to the main questions of this dissertation. Moreover, it summarizes the main findings and contributions of this theses whereas the second part pf this section will tap into final recommendations in terms future research and work as well as what can be done for the its subject.

Conclusion

Most of the countries in the world are spending a lot of investments researching and trying to find innovative ways to improve their public transportation network in order to fulfill community's needs, reduce traffic jams and congestions, stimulate the economy and boost attractiveness for businesses and foreign investments. Generally, most of the countries have found out that in order to enhance road traffic they have to build a strong case for motorists to persuade them to abandon their private vehicles and encourage them to rely more on public transportation instead.

With this goal in mind, local governments and agencies across the world have invested a lot in introducing modern transportation modes such as Metro, Water Taxi, Tram, modern bus, Taxi , etc. as well as developing and adopting the latest technologies and services associated with them. However, despite of these big investments, most of the statistics and reports have showed an increasing decline in the number of public transport passengers as opposed to private transportation users.

Additionally, most of the public transportation development plans and programs have seen several challenges and failures. The security concern that national security agencies and authorities in different parts of the world have expressed in terms of their ability to track and trace public transportation passengers in some highly sensitive national security cases has become a major one. Another less critical challenge is more operational and driven from the fact that increasing number of public transportation passengers has to be accompanied with increasing investments and expenditures in terms of staff, facilities, infrastructural services and, of course, technologies so as to ensure a smooth running of daily operations and avoid any unpleasant congestions and chaos.

The cultural challenge in some countries where public transportation is considered humiliating, degrading and something for lower class people only cannot be overlooked as well. Obviously, the use of public transportation varies from culture to another and from nation to another.

To overcome all these challenges and many others, concerned government authorities need to be more smart and think about this dilemma in a more holistic and strategic way. In doing so, they started to realize that need to leverage on the long-ignored and under-utilized resource they have had in their hands, which is Data.

Indeed, data has increasingly become one of the most valuable and strategic assets for all public and private organizations in today's world. Accordingly, this thesis comes to a conclusion that without utilizing and harnessing big data, transportation investments and plans will never realize their full potential and will never last long in the face of the ever-growing and emerging challenges.

This thesis has made effort trying to go beyond the traditional ways of analyzing available data. It went much farther than tracking and storing data to study and analyze data for the purpose of predicting and planning for the future in a more systematic, scientific and comprehensive manner.

The fundamental idea is not only about processing available data, it is about approaching and sourcing data before it comes. This means generating currently unavailable data using the one we do have. This is exactly where data mining comes to play and where we have to thank data mining science for providing us with the capabilities and features needed to predict future and be better prepared for the unforeseeable circumstances.

This theses supports that data and information extracted by robust data mining processes and technologies help public transportation authorities to mitigate and face their current and future challenges in a more smart and balanced way. To a larger extent, they can also enhance and improve their policy making and regulations setting processes based on the knowledge and the information they now have because of data mining and big data. It is worth mentioning though that using data mining will never lead to 100 percent trusted data, however it will provide much more accurate and reliable data than whatever will be generated from the transitional way of randomly using inconsistent and incomplete data and information.

As a matter of fact, successful strategic planning will never be complete and relevant without proper data mining processes in place. Increasingly, Strategy Departments are badly calling for more investments and enhancements in data mining in order to be more able to design effective strategies, consult on policies and regulations and also play a key role in supporting their deployment and improving their performance based on an accurate and ongoing measurement and analysis of the actual implementation.

As a more specific example to the case in hand, demographic data of public transportation passengers has been a focal point of this thesis; simply because demographic data provide an insight into the passengers' behaviors, patterns and characteristics such as age, gender, health status, car ownership, travel budget, trip distance, etc. all these provide wide possibilities for further analysis and exploration which will lead to effective regulations and policies as well as efficient and impactful transportation infrastructure investment decisions.

The equation here is very simple, the more personalized data you know about your customers, the better service you can provide to them and hence the more revenues you can generate and savings you can make.

In the transportation world, the story is very similar, as the more personalized data we know about daily passengers, the better service, the better security, the better operation and the more revenues can be attained. But, the question we may ask here is how we can know access all these information and how we can collect reliable demographic data of anonymous public passengers.

The answer is definitely through using effective data mining supported by right policies and effective programs and algorithms which will all lead to the creation of a massive platform that will be able to host all these data and make best use of them. In a more practical application, the transportation organization has managed to obtain accurate data about the gender of the passenger by analyzing their behavior. For example, passengers ending their journey at metro station A were 94% males which proved to be much more accurate than the old 28% figure that was available before adopting big data.

After going into these, this theses will add contributions on the following:

1. It is the first scientific thesis that analyze the public transportation on Middle East region. Most of the researchers that mining the transit network that took places on East Asia (China,

Korai, and Japan), Australia, Belgium, Netherlands and United States of America. Hence, this would a good start of all the researchers to start invest of the data got collected from the transportation mode in our reigns and countries.

2. Analyze the data that came from more that transportation mode, using data mining to predict the behavior of the passengers is still immature and requires a lot of data mining researchers. However, the result proofed that data-mining could be used to deliver a very important and valuable information to the business that were hiding.

3. This study was conducted with the significant need of the thesis analysis from the strategic planning department. When the result was ready, the result was discussing with the strategic department for their consultation and feedback.

4. With this research, we managed to increase the percentage of random figures for the demographic information for public passengers, and this will add a significant value in terms of national security, finical and strategic goals of public transportation adopting targets.

Future Work

As a future work, there is a huge potential to enhance the accuracy of this research by expanding the scope. I believe that increasing the sample collection to cover most of the circumstances would offer more data accuracy, as this research is only focusing on only one month.

Another future work, is to include more than two transportation mode and analyze the data from all the modes, such as Taxi, Water Taxi, Bus, Water Bus, Metro, Tram and even parking machines.

After reaching to this point, I think we can start thinking of integrating this information with another system like traffic system that provides more information about the public bus, and by proper analysis, we can answer the below question:

How many of private car owner using the public transportation modes? And how frequent? And what is the reason. After that, we can provide some information that can help on achieving the strategic gold converting resident to use the public transportation instead of private vehicles.

References

Shaw, S. L., & Xin, X. (2003). Integrated land use and transportation interaction: a temporal GIS exploratory data analysis approach. Journal of transport geography, 11(2), 103-115.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data.Knowledge and Data Engineering, IEEE Transactions on, 26(1), 97-107.

Zhan-Zhong, W., Feng-Hua, J., & Liu-Qing, Y. (2010). Data exhibition and mining for multimodal transport based on data warehouse. In 2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM) (Vol. 1, pp. 259-263).

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly, 36(4), 1165-1188.

Fiosina, J., Fiosins, M., & Müller, J. P. (2013). Big Data Processing and Mining for Next Generation Intelligent Transportation Systems. Jurnal Teknologi, 63(3).

Rahman, F. A., Desa, M. I., & Wibowo, A. (2011, June). A review of KDD-data mining framework and its application in logistics and transportation. InNetworked Computing and Advanced Information Management (NCM), 2011 7th International Conference on (pp. 175-180). IEEE.

Yuan, S., Liu, Y., Wang, G., Sun, H., & Zhang, H. A dynamic-data-driven driving variability modeling and simulation for emergency evacuation.

Wang, Z., Tu, L., Guo, Z., Yang, L. T., & Huang, B. (2014). Analysis of user behaviors by mining large network data sets. Future Generation Computer Systems, 37, 429-437.

Park, H. J., Kim, P. H., Marsico, M., & Rasheed, N. (2014). Data Mining Strategies for Real-time Control in New York City. Procedia Computer Science,32, 109-116.

Ge, G., & Tianyong, W. (2006, October). Data Mining Technique of Car Tracking in Transportation Service System Based on NN-FR. In Service Systems and Service Management, 2006 International Conference on (Vol. 1, pp. 133-137). IEEE.

Haluzová, P. (2008). Effective data mining for a transportation information system. Acta Polytechnica, 48(1).

Cheng, W., Ji, X., Han, C., & Xi, J. (2010, May). The Mining Method of the Road Traffic Illegal Data Based on Rough Sets and Association Rules. InIntelligent Computation Technology and Automation (ICICTA), 2010 International Conference on (Vol. 3, pp. 856-859). IEEE.

Luo, Q. (2008, May). Transportation Data Analyzing by Using Data Mining Method. In Information Processing (ISIP), 2008 International Symposiums on(pp. 766-767). IEEE.

Vogel, P., Greiser, T., & Mattfeld, D. C. (2011). Understanding bike-sharing systems using data mining: Exploring activity patterns. Procedia-Social and Behavioral Sciences, 20, 514-523.

Martín, L., Baena, L., Garach, L., López, G., & de Oña, J. (2014). Using Data Mining Techniques to Road Safety Improvement in Spanish Roads. Procedia-Social and Behavioral Sciences, 160, 607-614.

Kusakabe, T., & Asakura, Y. (2014). Behavioural data mining of transit smart card data: a data fusion approach. Transportation Research Part C: Emerging Technologies, 46, 179-191.

Dang, S., Hong, Z., Yang, S., & Baker, L. (2014, April). Intelligent urban traffic management system based on the energy-circle cards platform. In Information Science, Electronics and Electrical Engineering (ISEEE), 2014 International Conference on (Vol. 1, pp. 457-459). IEEE.

Ma, X., Wu, Y. J., Wang, Y., Chen, F., & Liu, J. (2013). Mining smart card data for transit riders' travel patterns. Transportation Research Part C: Emerging Technologies, 36, 1-12.

Cunningham, R. F. (1993, June). Smart card applications in integrated transit fare, parking fee and automated toll payment systems-the MAPS concept. InTelesystems Conference, 1993.'Commercial Applications and Dual-Use Technology', Conference Proceedings., National (pp. 21-25). IEEE.

Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies, 19(4), 557-568.

Husemann, D. (1999). The smart card: don't leave home without it.Concurrency, IEEE, 7(2), 24-27.

Jethi, A., Kalra, M., & Bhattacharyya, N. (2015). Analysis of Breast Cancer Recurrence using Combination of Data Mining Techniques.

bin Othman, M. F., & Yau, T. M. S. (2007, January). Comparison of different classification techniques using WEKA for breast cancer. In 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006 (pp. 520-523). Springer Berlin Heidelberg.

Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. Bioinformatics, 20(15), 2479-2481.

Stevanovic, D., An, A., & Vlajic, N. (2012). Feature evaluation for web crawler detection with data mining techniques. Expert Systems with Applications,39(10), 8707-8717.

Sharma, R., Alam, M. A., & Rani, A. (2012). K-means clustering in spatial data mining using weka interface. In International conference on advances in communication and computing technologies (ICACACT) (Vol. 26, p. 30).

Engel, T. A., Charão, A. S., Kirsch-Pinheiro, M., & Steffenel, L. A. (2014). Performance improvement of data mining in Weka through GPU acceleration.procedia computer science, 32, 93-100.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.