ASSOCIATION RULES MINING IN PLASTIC PIPE SECTOR

Nevzat Demir
Fırat Plastik Corporation, Istanbul, Turkey

n.demir@firat.com

Abstarct

In this study, an application has been made with data of plastic pipe sector within the data mining method. It has been tried to research whether there is an association rule between products sold by benefiting from sale data in this sector. Within this scope, the method of Association Rules mining has been used and analyses have been made on existing data with Apriori Algorithm of this model. As the result of these analyses, it has been detected that there are various association rules between the products sold.

Keywords: Data Mining, Association Rules, Data Storage, Apriori Algorithm, Plastic Pipe Sector

1. Introduction

Many institutions store their data in their servers; however, they cannot use these data out of the standard daily processes. Whereas, there may be many hidden patterns and relations that can direct the future of that institution within this data mass. Information that may be produced from these data could help in strategic decision making of the top management of the institution.

Institutions may offer special campaigns to their customers or distributors to make them more loyal and permanent. So, it is important to know the characteristics of these customers and detect their areas of interest. Scientific studies are made in the world for it. Data mining leads these studies. Data mining is a technique that can produce useful and usable information from these data.

This study consists of four parts. In the first part, descriptive information have been given in relation with data mining, association rules mining and Apriori algorithm used in the analysis. In the second part, literature samples made in relation with this area have been shown. In the third part, data storage have been prepared by using data of plastic pipe sector and association rules have been tried to be detected by making analysis. In the last part, findings have been assessed and accordingly, results have been interpreted.

2. Data Mining, Association Rules, and Apriori Algorithm

Data mining is known as a knowledge discovery process of analyzing data from different point of views. It involves using the data to work out into useful information which can be applied in various application, including advertisement, bioinformatics, database marketing, fraud detection, e-commerce, health care, security, web, financial forecasting etc(Jain et al:2001).

According to the Gartner Group, “Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques” (Larose 2005:XI).

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science (Han and Kamber, 2006:29).

The most common data mining algorithms and models, which include decision trees, associations, clustering, classification, multiple linear regression, sequential patterns, and timeseries forecasting, have the potential to identify drought patterns and characteristics. Association, clustering, and sequence discovery approaches may be useful tools for investigating and describing the occurrence and intensity of drought, while classification, regression, and time-series analyses may be appropriate for mapping and monitoring drought patterns.(Tadesse 2009).

There are many data mining algorithms, such as association rules, clustering, decision trees, discriminant analysis, artificial neural networks, genetic algorithms, and so on. These algorithms are used to process data from various fields to retrieve information and discover knowledge that can drive an executive’s decisions. Information is data associated with the past and the present. Knowledge provides a basis for the prediction of future trends based on original data and the necessary information extracted from the original data. Clearly, information and knowledge are communicated through data (Wu and Li 2003).

Association rules mining is one of the commonly used type of data mining. In this method which is also known as “Association Rule”, “Association Analysis” and “Market Basket Analysis” the main focus of research is of the consumption habits of

the customers. In consequence, decisions are made concerning the question of ”which product should be placed near which product on the shelves of the stores.” For example, the main goal is to create a perception that will lead the customer to purchase the other product while purchasing one (Alan 2014).

The goal of association rules is to detect relationships or associations between specific values of categorical variables in large data sets. This technique allows analysts and researchers to uncover hidden patterns in large data sets (Nisbet et al,2009).

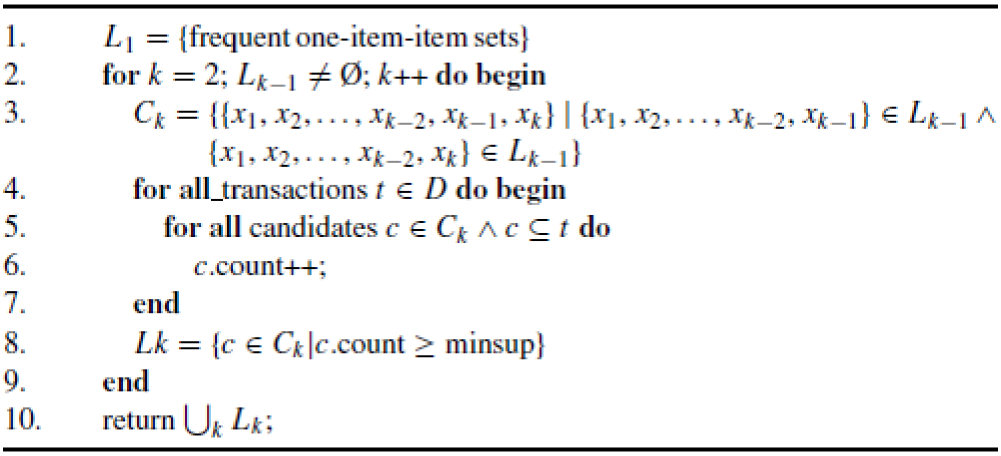
Association rules were developed in the field of computer science and are often used in important applications such as market basket analysis, to measure the associations between products purchased by a particular consumer, and web click stream analysis, to measure the associations between pages viewed sequentially by a website visitor. In general, the objective is to underline groups of items that typically occur together in a set of transactions (Guidici et al,2009).

One of the algorithms commonly used in association rules is Apriori algorithm.

Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules. It explores the level-wise mining Apriori property that all nonempty subsets of a frequent itemset must also be frequent. The name of the algorithm is based on the fact that the algorithm uses *prior knowledge* of frequent itemset properties, as we shall see following.
Apriori employs an iterative approach known as a *level-wise* search, where *k*-itemsets are used to explore (*k*+1) itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted *L*1. Next, *L*1 is used to find *L*2, the set of frequent 2-itemsets, which is used to find *L*3, and so on, until no more frequent *k*-itemsets can be found. The finding of each *Lk* requires one full scan of the database. To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property, presented below, is used to reduce the search space. We will first describe this property, and then show an example illustrating its use (Han and Kamber, 2006:234-235).

The *Apriori* algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size *k*+1 are created by joining a pair of frequent itemsets of size *k* (this is known as the candidate generation step). A candidate is discarded if any one of its subsets is found to be infrequent during the candidate pruning step (Tan el al,2006:79).

General structure of Apriori Algorithm is as follows (Webb 2003):



```
$L_1 = \{\text{frequent one-item-item sets}\}$
for $k = 2; L_{k-1} \neq \emptyset; k{+}{+}$ do begin
    $C_k = \{\{x_1, x_2, \ldots, x_{k-2}, x_{k-1}, x_k\} \mid \{x_1, x_2, \ldots, x_{k-2}, x_{k-1}\} \in L_{k-1} \wedge$
        $\{x_1, x_2, \ldots, x_{k-2}, x_k\} \in L_{k-1}\}$
    for all_transactions $t \in D$ do begin
        for all candidates $c \in C_k \wedge c \subseteq t$ do
            $c$.count++;
    end
    $Lk = \{c \in C_k | c.\text{count} \geq \text{minsup}\}$
end
return $\bigcup_k L_k$;
```

3.Summary of Literature

Many researches have been made in the literature with association rules mining technique.

Kumar and Rukmani (2010) have produced association rules from web log files by using Apriori and FP-Growth algorithms. Alan (2014) has detected by using student data that there are association rules between both succeeded courses and failed lessons. Babu and Bhuvaneswari (2012) have applied Association Rules Mining to Customer Relations Management by using data of customers of a company. Umarani and Punithavalli (2011) have made analysis with different association rules mining algorithms on real life data such as retail sale data and market basket data. Erpolat (2012) has made research on association rules between service equipment bought by customers in an automobile service. Wang et al. (2010), have applied Fuzzy FP-Growth algorithm and standard Apriori algorithm on two different datasets and found that Fuzzy association rules display better performance on both cases. Ivancsy et al. (2005), have compared the advantages and disadvantages of the most popular algorithms of association rules. According to their findings, Apriori and FP-Growth algorithms on semantic datasets generated by dataset generator downloaded from IBM website and propounded those cubic algorithms of both algorithms display better performance.

4. Preparation of Data

In this study, sale information obtained from the PVC sector have been used. Data have been provided in Excel format from a firm named *Fırat Plastic Corporation Türkoba Quarter Fırat Plastic Avenue*. Data storage have been prepared by using Excel macros.

Data warehouse is a subject-oriented, integrated, time-variant, non-updatable collection of data used in support of management decision-making. A data warehouse is the data and the process managers that make information available, enabling people to make informed decisions. Before the use of data warehouse, companies used to store data in separate databases, each of which were meant for different functions (Bose et al,2009).

A data ***warehouse*** is a database system used to store information from various operational databases for decision support purposes. A data warehouse for a retailer might include information from a market basket database, a supplier database, customer databases, etc. The data in the payroll database might not be in the data warehouse if they are not considered to be crucial in decision support. A data warehouse is not created just by dumping the data from various databases to a single disk. Several integration tasks have to be carried out, such as resolving possible inconsistencies between attribute names and usages, finding out the semantics of attributes and values, and so on. Building data warehouses is often an expensive operation, as it requires much manual intervention and a detailed understanding of the operational databases (Hand, et al,2001:419)

The construction of a data warehouse requires data cleaning, data integration, and data consolidation. The utilization of a data warehouse often necessitates a collection of *decision support* Technologies (Han and Kamber, 2006:107).

In this study, required conversions have been made upon data and written on the text file named “Veriset.arff”. In this conversion stage, noisy data have been sorted and prevented from being written on the text file. During preparation of data storage, the value “1” has been assigned in case of purchase of product. In the cases where no information is found for purchase of the product, the value “?” has been assigned.

5. Application

In this study, an application has been made with data of plastic pipe sector within the data mining method. It has been tried to research whether there is an association rule between products sold by benefiting from sale data.

In the study carried out, 3.6.11 version of WEKA Program (Waikato Environment for Knowledge Analysis) having developed in Waikato University has been used. WEKA Program is open-source software. This program supports many algorithms of classification, clustering, and association rules. It supports WEKA, text-based arff, arff.gz, names, data, csv, c45, libsvm, dat, bsi, xrff, xrff.gz file types as well as URL addresses in which there are data bases and data.

Data consisting of 893.656 rows and four columns (date, customer, product, and amount) have been used as input from the sale reports of 2010-2012 obtained from the firm of *Fırat Plastik Corporation Türkoba Quarter Fırat Plastik Avenue* operating İstanbul. By taking customer having bought minimum 5 units and product having been sold minimum 10 units among these data into consideration, association rules mining have been made between 2543 sale and 137 different products. As the result of the analysis carried out, 24 rules have been produced with Apriori algorithm.

24 rules found with **Apriori** application by existing data are as follows.

```
1. 110/45 ACIK DİRSEK TSE275=1 4 ==> 100/250  DUBLEX 3.2 MM ATIK SU
BORU=1 4    conf:(1) lift:(158.88) lev:(0) [3]
 2. 100/3000 GEDİZ PVC BORU=1 100/500 GEDİZ PVC BORU=1 3 ==> 70/3000 GEDİZ
PVC BORU=1 3 conf:(1) lift:(317.75) lev:(0) [2]
 3. 100/3000 GEDİZ PVC BORU=1 70/3000 GEDİZ PVC BORU=1 3 ==> 100/500 GEDİZ
PVC BORU=1 3    conf:(1) lift:(423.67) lev:(0) [2]
 4. 100/500 GEDİZ PVC BORU=1 70/3000 GEDİZ PVC BORU=1 3 ==> 100/3000 GEDİZ
PVC BORU=1 3    conf:(1) lift:(282.44) lev:(0) [2]
 5. 25 FIRATTHERM SIVA ALTI VANA=1 4 ==> 20 FIRATTHERM SIVA ALTI
VANA=1 3    conf:(0.75) lift:(272.36) lev:(0) [2]
 6. 50/2000 GEDİZ PVC BORU=1 4 ==> 50/500 GEDİZ PVC BORU=1 3    conf:(0.75)
lift:(381.3) lev:(0) [2]
 7. 70/1000 GEDİZ PVC BORU=1 6 ==> 100/150  DUBLEX 3.2 MM ATIK SU BORU=1
4    conf:(0.67) lift:(112.98) lev:(0) [3]
 8. 50/500 GEDİZ PVC BORU=1 5 ==> 50/2000 GEDİZ PVC BORU=1 3    conf:(0.6)
lift:(381.3) lev:(0) [2]
 9. 100/500 GEDİZ PVC BORU=1 6 ==> 100/3000 GEDİZ PVC BORU=1 3    conf:(0.5)
lift:(141.22) lev:(0) [2]
```

10. 100/500 GEDİZ PVC BORU=1 6 ==> 70/3000 GEDİZ PVC BORU=1 3   conf:(0.5) lift:(158.88) lev:(0) [2]

11. 100/500 GEDİZ PVC BORU=1 6 ==> 100/3000 GEDİZ PVC BORU=1 70/3000 GEDİZ PVC BORU=1 3   conf:(0.5) lift:(423.67) lev:(0) [2]

12. 20 FIRATTHERM SIVA ALTI VANA=1 7 ==> 25 FIRATTHERM SIVA ALTI VANA=1 3   conf:(0.43) lift:(272.36) lev:(0) [2]

13. 70/3000 GEDİZ PVC BORU=1 8 ==> 100/3000 GEDİZ PVC BORU=1 3   conf:(0.38) lift:(105.92) lev:(0) [2]

14. 20 MM FIRATTHERM ’BEYAZ’ T-PARÇA=1 8 ==> 100/500 DUBLEX 3.2 MM ATIK SU BORU=1 3   conf:(0.38) lift:(105.92) lev:(0) [2]

15. 70/3000 GEDİZ PVC BORU=1 8 ==> 100/500 GEDİZ PVC BORU=1 3   conf:(0.38) lift:(158.88) lev:(0) [2]

16. 70/3000 GEDİZ PVC BORU=1 8 ==> 100/3000 GEDİZ PVC BORU=1 100/500 GEDİZ PVC BORU=1 3   conf:(0.38) lift:(317.75) lev:(0) [2]

17. 100/500 DUBLEX 3.2 MM ATIK SU BORU=1 9 ==> 100/250 DUBLEX 3.2 MM ATIK SU BORU=1 3   conf:(0.33) lift:(52.96) lev:(0) [2]

18. 100/3000 GEDİZ PVC BORU=1 9 ==> 100/500 GEDİZ PVC BORU=1 3   conf:(0.33) lift:(141.22) lev:(0) [2]

19. 100/3000 GEDİZ PVC BORU=1 9 ==> 70/3000 GEDİZ PVC BORU=1 3   conf:(0.33) lift:(105.92) lev:(0) [2]

20. 100/500 DUBLEX 3.2 MM ATIK SU BORU=1 9 ==> 20 MM FIRATTHERM ’BEYAZ’ T-PARÇA=1 3   conf:(0.33) lift:(105.92) lev:(0) [2]

21. 100/3000 GEDİZ PVC BORU=1 9 ==> 100/500 GEDİZ PVC BORU=1 70/3000 GEDİZ PVC BORU=1 3   conf:(0.33) lift:(282.44) lev:(0) [2]

22. 100/150 DUBLEX 3.2 MM ATIK SU BORU=1 15 ==> 70/1000 GEDİZ PVC BORU=1 4   conf:(0.27) lift:(112.98) lev:(0) [3]

23. 100/250 DUBLEX 3.2 MM ATIK SU BORU=1 16 ==> 110/45 ACIK DİRSEK TSE275=1 4   conf:(0.25) lift:(158.88) lev:(0) [3]

24. 100/250 DUBLEX 3.2 MM ATIK SU BORU=1 16 ==> 100/500 DUBLEX 3.2 MM ATIK SU BORU=1 3   conf:(0.19) lift:(52.96) lev:(0) [2]

The two statistics that were used initially to describe these relationships were support and confidence. These are numeric values. To describe them we need to define some numeric terms. Let D be the database of transactions and N be the number of transactions in D. Each transaction Di is an item set. Let support(X) be the proportion of transactions that contain item set X:

$$support(X) = \left|\{I \mid I \in D \wedge I \supseteq X\}\right| / N$$

where I is an item set and $|.|$ denotes the cardinality of a set.

The support of an association rule is the proportion of transactions that contain both the antecedent and the consequent. The confidence of an association rule is the proportion of transactions containing the antecedent that also contain the consequent. For an association

A →C,

$$support(A \rightarrow C) = support(A \cup C)$$

$$confidence(A \rightarrow C) = support(A \cup C)/support(A).$$

If support is sufficiently high (and the transactions represent a random sample from the same data distribution as future transactions), then confidence is a reasonable estimate of the probability that any given future transaction that contains the antecedent will also contain the consequent (Webb 2003).

The most popular objective measure of interestingness is lift.

$$Lift(A \rightarrow C) = confidence(A \rightarrow C)/support(C).$$

This is the ratio of the frequency of the consequent in the transactions that contain the antecedent over the frequency of the consequent in the data as a whole. Lift values greater than 1 indicate that the consequent is more frequent in transactions containing the antecedent than in transactions that do not (Webb 2003).

A measure that captures both the volume and the strength of the effect in a single value is leverage.

$$Leverage(A \rightarrow C) = support(A \rightarrow C) - support(A) \times support(C)$$

This is the difference between the observed frequency with which the antecedent and consequent co-occur and the frequency that would be expected if the two were independent (Webb 2003).

6. Conclusion and Assessment

This technique which is one of the most commonly used types of data mining researches whether association rule occurs in different areas. With this method, it may help customers by reminding the other products that they may probably buy as they buy any product as well as it ensures increase in sales by allowing businesses grouping of products between which there is association rule. It has been tried to be researched in this study that whether there is association rule between the sale data in the sector of plastic pipe and, if so, reliability level of this rule is high by making analysis on data of this sector by the method of Association Rules which is one of the data mining techniques. The level of reliability level shows the level of association rule.

24 rules have been able to be produced as the result of the analysis made. When produced rules are taken into consideration, it is understood in the first rule that customers having bought 110/45 ACIK DİRSEK TSE275 buy 100/250 DUBLEX 3.2 MM ATIK SU BORUSU with the reliability level of 1. It is revealed in the second rule that those having bought 100/3000 GEDİZ PVC BORU and 100/500 GEDİZ PVC BORU buy 70/3000 GEDİZ PVC BORU with the reliability level of 1. In the

third rule, it has been detected that those buying 100/3000 GEDİZ PVC BORU and 70/3000 GEDİZ PVC BORU buy 100/500 GEDİZ PVC BORU with the reliability level of 1. Similar assessments have been made for the remaining 21 rules.

As the result of these analyses, it has been asserted that there are association rules between the products bought or sold in the plastic pipe sector. Businesses working on plastic pipe product sales may present products in their stores in a way that shall remind them to customers by taking these rules into consideration. This condition shall both increase the profitability of the business and ease the work of customer. At the same time, it shall provide time saving for business and customer, increase customer satisfaction, and loyalty of customer for the business.

7. REFERENCES

1. Alan, MA, (2014), Association rules mining: An analysis on student grades, **Technics Technologies Education Management-TTEM**, Vol:9, No:1, pp:172-178

2. Babu G. and Bhuvaneswari, T. (2012), “A Data Mining Technique To Find Optimal Customers For Beneficial Customer Relationship Management”, *Journal of Computer Science* 8(1): 89-98.

3. Bose, Indranil,&. Chun, Lam Albert Kar,& Yue, Leung Vivien Wai,& Ines, & Wan, Li Hoi, & Helen, Wong Oi Ling**, Business Data Warehouse: The Case of Wal-Mart,** Data Mining Applications for Empowering Knowledge Societies, pp.190, Ed. Hakikur Rahman, Information Science Reference, pp.190,

4. Erpolat, Semra (2012), “Comparison of Apriori and FP-Growth Algorithms in determining Association Rules in Automobile Authorized Services, 12(2):137-146.

5. Giudici, Paolo,& Figini, Silvia, (2009) **Applied Data Mining For Business and Industry**,Second Edition, Wiley Publication, West Sussex, pp:90-91

6. Han, Jiawei & Kamber, Micheline, (2006) **Data Mining: Concepts and Techniques**, *Second Edition*, Morgan Kaufmann Publications, p:234-235, San Francisco

7. Hand David & Mannila, Heikki & Smyth, Padhraic, **(2001), Principles of Data Mining, ,** e-book, MIT Press, Cambridge pp.419

8. Ivancsy R, Vajk I. (2005), “Fast Discovery Of Frequent Itemsets: A Cubic Structure-Based Approach”. Informatica , 29, pp:71–78.

9. Jain, Yogendra Kumar & Yadav, Vinod Kumar& Panday, Geetika S., (2011). **An Efficient Association Rule Hiding Algorithm For Privacy Preserving Data Mining**, International Journal On Computer Science And Engineering, Vol. 3 No. 7, ISSN : 0975-3397, pp. 2792-2798

10. Kumar, B. Santhosh and Rukmani, K. V. (2010), "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms", International Journal of Advanced Networking and Applications, 1(6): 400-404.

11. Larose, Daniel T., (2005). **Discovering Knowledge In Data**, Wiley Publication, pp. xi, New Jersey

12. Nisbet, Robert, & Elder, John, & Miner, Gary, (2009). **Handbook Of Statistical Analysis And Data Mining Applications**, Elsevier Inc, Burlington, pp.126

13. Tadesse, T., Wardlow, B. And Hayes, M.J. (2009), **"**The Application of Data Mining for Drought Monitoring and Prediction", **Data Mining Applications for Empowering Knowledge Societies,** Ed. Hakikur Rahman, Hershey, New York, pp.280-291.

14. Tan PN, Steinbach M, Kumar V. (2006),Introduction to Data Mining, Pearson Addison-Wesley.

15. Umarani, V. and Punithavalli, M. (2011), "An Empirical Analysis Over The Four Different Methods of Progressive Sampling-Based Association Rule Mining" European Journal of Scientific Research 66(4): 620-630.

16. Wang C, Lee W, Pang C.(2010), "Applying Fuzzy FPGrowth to Mine Fuzzy Association Rules". World Academy of Science, Engineering and Technology, 2010; 65: 956-962.

17. Webb, G.,I. (2003). Association Rules. In Nong Ye (Edt.), **The Handbook Of data Mining**, pp. 27-28,. New Jersey: Lawrence Erlbaum Associates,Inc.

18. Wu, Tong and Li, Xiangyang (2003), "**Data Storage and Management**", Nong Ye (ed.), The Handbook of Data Mining, New Jersey: Lawrence Erlbaum Associates Inc., 393-407.