I.J. Information Technology and Computer Science, **2019, 10, 34-39**
Published Online October 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2019.10.05

An Association Prediction Model: GECOL as a Case Study

Ashraf Mohammed Abusida
Institute of Science and Technology, Kastamonu University, Kastamonu, 37150, Turkey
E-mail: aabusida@ogr.kastamonu.edu.tr

Yasemin Gültepe
Department of Computer Engineering, Facult of Engineering and Architecture, Kastamonu University, Kastamonu, 37150, Turkey
E-mail: yasemingultepe@kastamonu.edu.tr

Received: 17 April 2019; Accepted: 23 May 2019; Published: 08 October 2019

***Abstract*—**Nowadays, there exists a lot of information that can be handled from business transactions and scientific data and information retrieval is simply no longer enough for decision-making. In this paper will supervised machine learning technique is applied to the mine data warehouse for Enterprise Resource Planning (ERP) of the General Electricity Company of Libya (GECOL). This technique has been applied for the first time on the data of production, transportation and distribution departments. These data are in the form of purchase and work orders of operational material strategic equipment spare parts. This technique would extract prediction rules in order to assist the decision-makers of the company to make appropriate future decisions more easily and in less time. A supervised machine learning technique has been adopted and applied for the mining data warehouse. A well-known software package for data mining which is referred to as WEKA tool was adopted throughout this work. The WEKA tool is applied to the collected data from GECOL. The conducted experiments produce prediction models in the form set of rules in order to help responsible employees make the suitable, right and accurate future decision in a simple way and inappropriate time. The collected data were preprocessed to be prepared in a suitable format to be fed to the WEKA system. A set of experiments has been conducted on those data to obtain prediction models. These models are in the form of decision rules. The produced models were evaluated in terms of accuracy and production time. It can be concluded that the obtained results are very promising and encouraging.

***Index Terms*—**Machine learning, Data mining, Enterprise resource planning, Data warehouse.

I. Introduction

The problem is addressing arose from the presence of a huge amount of data stored in the database of General Electricity Company of Libya (GECOL) and represented in the database of Enterprise Resource Planning (ERP) [1,2]. The database is in the form of purchase orders and requests for spare parts of equipment. The equipment is strategic such as (electrical generators, transformers of high-voltage, towers transport, distribution networks, and control devices) or equipment service assistance such as (auxiliary equipment). The decision makers of GECOL are not taking advantages of these data in the development or extraction rules to help decision-makers in making strategic decisions as quickly as required by the maintenance phase in both power station and networks of high-voltage transmission or distribution stations. Accordingly, the idea of this paper is to invest and analyze the available digital format data in producing beneficial prediction models in terms of production rules that in turn help decision makers of the GECOL to make appropriate the right and quick decision. The obtained knowledge would provide the decision-makers with the essential information that can be used to draw future plans for development projects.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 provides information on data mining and machine learning techniques. Section 4 provides information about GECOL data set and results of association prediction models (Distribution Site (D01), Production Site (G01) and Transportation Site (T01)). Finally, Section 5 presents conclusions.

II. Related Works

Several relevant studies of curricula related to machine learning and data mining technique have been reviewed. Since the early 2003s, many researches have been done on data mining technique, and extracting rule form dataset. All relevant studies have provided us clearly on the technique and applying machine learning algorithms, but none of them focus on the mine ERP data for GECOL and implement some of the measures in terms of preparing and dividing the raw data and applying machine learning technique through specific algorithm and identify some of the factors that are working on this technique to get the best results, thus producing

prediction models.

Nowadays, applied data mining techniques [3] are widely used to discover a new and comprehensive data set. The data mining process generates several patterns from a given data source. The most well-known data mining tasks are the process of discovering frequent itemsets, frequent sequential patterns, frequent sequential rules, and frequent association rules. Numerous efficient algorithms have been proposed to perform the above operations [4].

In Zhou and Wang (2010) a rough set is provided to establish the association rules used in diagnosing the power transformer. The rough set can establish deep relationship, the power transformer association rule is won by the rough set. By reducing the coarse set, the substitute feature that affects the grading performance is deleted. Then the power transformer association rule is acquired. Experimental results show that the method has very good results [5].

Association rules are used successfully in determining the observed features together. In Ivančević et al. (2015) study, the association between ECC (Early Childhood Caries) related factors in children was evaluated and dominant risk factors were analyzed by using merger rule, one of the data mining techniques. As a result of this study, male sex, frequent breastfeeding (with risk factors), high birth order, tongue and low body weight at birth were found to be the dominant risk factors. However, parents' low recovery sensitivity was significantly associated with ECC only in boys [6].

Transformer condition assessment system is established based on data mining for a routine/preventive test. Also, the comprehensive analysis is used to classify for the independent status parameters. Synthetic status parameters are taken as the key elements of the transformer condition assessment, which is conducive to an accurate assessment. The constant weight coefficients of status parameters are calculated through association rules to avoid over-reliance on expert opinion or subjective experience, and to reflect the weight of each status parameter based on objective facts. The variable weight coefficients are used to calculate the condition score of power transformers [7].

Every year, traffic accidents kill more than one million people and injure more than 20 million people worldwide. Daher et al. [8] aimed to provide guidance on road safety and to raise awareness by identifying the main causes of traffic accidents. In this study, the Frequency Pattern Growth algorithm has been used to improve the knowledge and to establish association rules to highlight the time and environment settings that cause the most devastating accidents.

While this concept is versatile and often open to interpretation, the term “Industry 4.0”" has a clear theme of intelligent manufacturing; it leverages advanced computational technologies and leverages advances in digital systems and machine learning processes to support decision-making, self-sufficient work through distributed control networks, and to self-correct and correct problems to reveal them. Oliff and Liu (2017) have demonstrated how data mining principles can be used by focusing on ways to integrate production paradigms into existing production processes, particularly on improving product and process quality to begin exploring the concept of Intelligent Manufacturing within Industry 4.0 [9].

Hierarchical clustering technique has been applied for customer clustering and advanced apriori algorithm for purchasing pattern analysis. Apriori algorithm [10] has been developed especially in data mining studies on very large scale databases. The results showed that the proposed approach is capable of clustering high-profit, high-standard, low-risk, high-focus and low-care customers. In addition, we can say that a cluster of high-profit has high revenue customers and the cluster of high focus represents target customer. The results also concluded that the rules generated by the proposed algorithm heighten the association mining of items set with proper placement [11].

Applied data mining techniques on manufacturing data are used to help the manufacturer obtain interesting and valuable information. In the study of Ismail et al., apriori algorithm was used to obtain the rules of unity and to predict the most common production. Also, they used the k-means algorithm to uncover the link between the observed frequent patterns [12].

III. Materials and Methods

A. Data Set

The dataset used for experiments in this paper was historical data on orders to purchase procedures and requests for equipment parts for the GECOL within a specified period of time. This work is performed on some data for analysis and study and thus finding the bonding rules, through a series of procedures and experiments as shown in Table 1.

Table 1. GECOL Dataset Properties

Variables	Unit	Description
ORDER_NO	String	Represent of work order number or request number
PART_NO	String	Spare parts code
PART_DESCRIPTION	String	Description of spare parts (Represents the full name of spare part)
SITE	String	Site is description to any sector follow this spare part (G01-Production, D01-Distribution, T01-Transportation)
UNIT	String	Unit of spare parts (m, m^3, ea, l, kg, bkl, set)

B. Data Mining

The information age will leads to power and success due to sophisticated technologies such as computers, satellites, etc. Tremendous amounts of information have been collected. Initially, with the advent of computers and means for mass digital storage, all sorts of data are collected and stored, counting on the power of computers to help sort through this amalgam of information.

Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming, and is not utilized adequately [13].

Data mining is a new and powerful technology with great potential to help companies focus on the most important information in their data warehouses, extracting confidential information from large databases. Data Mining studies algorithms and computational paradigms that allow computers to discover structure in databases, perform prediction and forecasting and generally improve their performance through interaction with data. Machine learning is concerned with building computer systems that have the ability to improve their performance in a given domain through experience.

A well-known software package for data mining is referred to as WEKA tool. WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand, in 1993, the University of Waikato in New Zealand started development of the original version of Weka [14]. The system provides a rich set of powerful Machine Learning algorithms for data mining tasks, some not found in commercial data mining systems. These include basic statistics and visualization tools, as well as tools for pre-processing, classification, and clustering, all available through an easy to use graphical user interface.

Machine learning and data mining are becoming increasingly important areas of engineering and computer science and have been successfully applied to a wide range of problems in science and engineering, therefore this technique has been used and adopted in this study.

The series of procedures and experiments included three stages, data selection, data preparation, applying the algorithm and strong rule extraction, as shown in Figure 1.

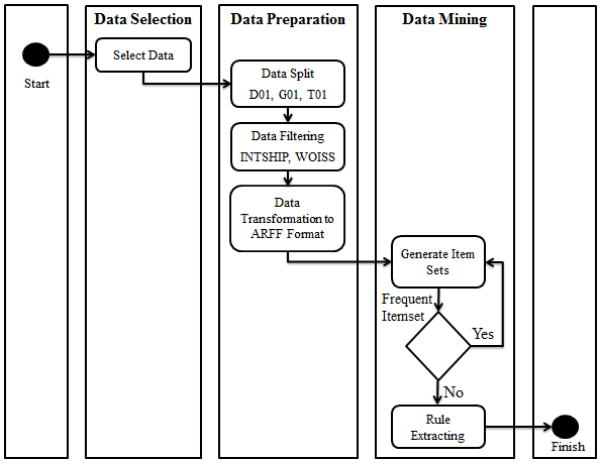

Fig.1. Process model for ERP of GECOL

To obtain accurate results, the historical (previous) transactions of the ERP system are used in our experimental work. A set of forms has been extracted in order to analyze the relationships amongst the distinct and various spare parts and partition them according to the work type. The spare part code has been used instead of spare part description to facilitate understanding of the resulting rules for technicians and engineers of GECOL. The spare parts are partitioned into three sets. These are all GECOL sectors which are (Distribution, Production and Transportation) as shown in Table 2.

Table 2. Sites

Site	Site's Symbol
Production	G01
Distribution	D01
Transportation	T01

This partitioning allows finding results much better than those found by combining all GECOL sectors together as one unit that affects negatively on the significance of materials. In addition, this portioning assists in forming a clear picture for each sector and its results will not influence the results of remaining sectors.

C. Association Prediction Model

Using WEKA program, data files were used that has been prepared and filtered and divided files for each site (D01, G01 and T01). The rows represent the purchase orders and requests work orders, the columns represent the spare parts code, and the letter "t" in the table indicates the existence or included of the spare parts by this order, as shown in Figure 2, and implementation of the apriori algorithm.

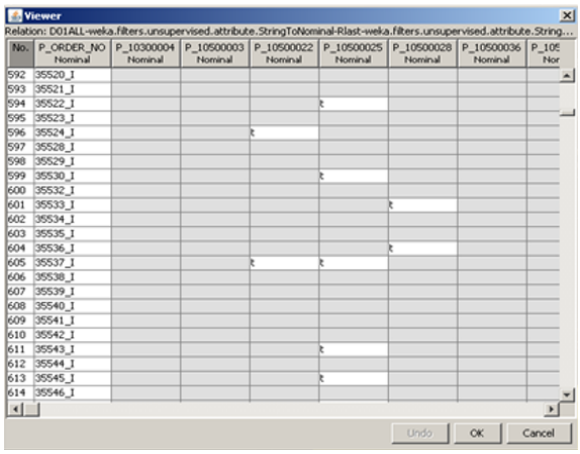

Fig.2. Data file for D01 site (D01.Arff)

IV. Experimental Analysis

A. Results of Distribution Site (D01)

Apriori algorithm has been applied on the data file "D01.Arff" produces the results relevant to the Distribution Site. It has been determined minimum support and minimum confidence for optimal results. A sample of the obtained results (the best rules found) is shown in Table 3. The produced rules are listed sequentially.The Association Rules find items that are above a certain Support value and generate the desired rules that belong to a certain Confidence among the remaining items thereafter [15].

Table 3. Best Rules for D01 Site

No	Rules	Confidence
1	P_10800046 and P_10800071 and P_10809316 and P_10809337 ==> P_10810978	100%
2	P_10800046 and P_10800071 and P_10809337 ==> P_10810978	99%
3	P_10800046 and P_10800071 and P_10809316 ==> P_10810978	99%
4	P_10600047 and P_10600071 ==> P_10600097	99%
5	P_10800046 and P_10800118 and P_10809337 ==> P_10810978	99%
6	P_10800046 and P_10800049 and P_10800118 and P_10809337 ==> P_10810978	99%
7	P_10500074 and P_10800046 and P_10800071 and P_10809337 ==> P_10810978	99%
8	P_10800046 and P_10800049 and P_10800071 and P_10809337 ==> P_10810978	99%
9	P_10800046 and P_10800071 and P_10800118 and P_10809337 ==> P_10810978	99%
10	P_10500074 and P_10800046 and P_10809316 ==> P_10810978	99%

Support: The ratio of the number of actions containing an asset to the total number of actions. The probability value is shown as the rate of the number of all calculations (Total) involving both X and Y to the number of all calculations in Equation 1.

$$Support = \frac{(X+Y)}{Total} \quad (1)$$

Confidence: The ratio of the number of actions involving two entities to one. The confidence value is a conditional probability criterion and is shown as the rate of the number of calculations involving both X and Y to the number of calculations involving only X in Equation 2.

$$Confidence = \frac{(X+Y)}{X} \quad (2)$$

The values of support and trust are sought to establish a partnership rule. The rule requires minimum support and minimum trust. With these results, the rules are found by applying the Equation 2 confidence formula.

B. Result of Production Site (G01)

Running the WEKA system (specifically applying the Apriori algorithm) on the data file "G01.Arff" produces the results relevant to the Generation site. A sample of the obtained results (the best rules found) is shown in Table 4 as listed sequentially.

Table 4. Best Rules for G01 Site

No	Rules	Confidence
1	P_11000299 and P_11801157 ==> P_11000444	100%
2	P_11000708 ==> P_11000706	97%
3	P_12500819 ==> P_12500832	75%
4	P_11000299 and P_11000444 ==> P_11801157	75%
5	P_12500832 ==> P_12500819	67%
6	P_10200014 ==> P_10200389	67%
7	P_11801157 ==> P_11000444	63%
8	P_11000444 and P_11801157 ==> P_11000299	60%
9	P_10200013 ==> P_10200293	57%
10	P_10200293 ==> P_10200013	44%

Going through the obtained results, it shows that the highest confidence between the spare parts (P_11000299 and P_11801157) and the spare part (P_11000444) is (100%), while the lowest value of confidence between spare parts (P_10200293 and P_10200013) is 44%, which are represented in rule 10.

C. Result of Transportation Site (T01)

Apriori algorithm has been applied on the data file "T01.Arff" produces the results relevant to the Transportation Site. It has been determined minimum support and minimum confidence for optimal results. A sample of the obtained results (the best rules found) is shown in Table 5. The produced rules are listed sequentially.

Base on the rule obtained, it shows that the highest confidence (100%) between (P_10802670 and P_10818313) in rule 1, while the lowest value of confidence (75%) between spare parts (P_10819957 and P_10802428) are shown in Rule 10.

Table 5. Best Rules for T01 Site

No	Rules	Confidence
1	P_10802670 ==> P_10818313	100%
2	P_10802670 ==> P_10819976	100%
3	P_10819969 ==> P_10810593	100%
4	P_10818313 ==> P_10819976	100%
5	P_10905207 ==> P_10905208	100%
6	P_10818313 and P_10819976 ==> P_10802670	100%
7	P_10802670 and P_10819976 ==> P_10818313	100%
8	P_10802670 and P_10818313 ==> P_10819976	100%
9	P_10820392 ==> P_10802739	89%
10	P_10819957 ==> P_10802428	75%

V. Conclusion

The rules that have been tested in accordance with the results were expected, which is the generation of new rules for each site, and achieve the main goals. In this paper, very interesting, beneficial, and significant results have been obtained.

It is clear that some rules have been obtained with a high confidence factor of up to 100%, as in Rule 1 for D01 Site Rule1: (P_10800046 and P_10800071 and P_10809316 and P_10809337 ==> P_10810978). Which means that the ratio of the requirement to request the spare part (P_10810978) up to 100% when creating a work order or purchase order includes spare parts (P_10800046, P_10800071, P_10809316, P_10809337) combined. Also, the result of Rule 3 of G01 Site indicated a confidence factor of up to 75% Rule3: (P_12500819 ==> P_12500832) Which means that the ratio of the requirement to request the spare part (P_12500832) up to 75% when the creation of the purchase order includes spare parts (P_12500819).

While the result of Rule 9 of T01 Site indicated a confidence factor of up to 89% Rule 9: (P_10820392 ==> P_10802739). Which means that confidence factor between spare part (P_10820392 and P_10802739) up to 89%. From our point of view, the results are very encouraging and promising.

Also, this paper conducted a set of experiments to build a model in terms of predictive rules, and the obtained predictive model has been evaluated. The obtained predictive model is very useful, it helps researchers, developers, engineers, and technicians in each sector to extract the key and necessary information. It also helps effectively in planning and improving services. It can be concluded that the obtained predictive model is a foundation for GECOL strategic information. In addition, it can be used to help responsible employees make their right, suitable, necessary, and quick decisions properly and unmistakably.

References

[1] W. Alsuessi, "General Electricity Company of Libya (GECOL), European International Journal of Science and Technology, Vol. 4 No. 1, 2015.

[2] D. M. Bahssas, A. M. AlBar and R. Hoque, "Enterprise Resource Planing (ERP) Systems: Design, Trends and Deployment", The International Technology Management Review, Vol. 5, No. 2, pp. 72-81, 2015.

[3] T. Slimani and A. Lazzez, "Efficient Analysis of Pattern and Association Rule Mining Approaches", International Journal of Information Technology and Computer Science, Vol. 6, No. 3, pp. 70-81, 2014.

[4] P. Giudici and S. Figini, "Applied Data Mining for Business and Industry", Second Edition, Wiley.

[5] M. Zhou and T. Wang, "Fault Diagnosis of Power Transformer Based on Association Rules Gained by Rough Set", The 2nd International Conference on Computer and Automation Engineering, 2010.

[6] V. Ivančević, I. Tušek, M. Knežević, S. Elheshk and I. Luković, "Using Association Rule Mining to Identify Risk Factors for Early Childhood Caries", Computer Methods And Programs In Biomedicine , vol. 122, pp. 175–181, 2015.

[7] L. Li, X. Longjun, Z. Deng, Y. Bin, G. Yafeng and L. Fuchang, "Condition Assessment of Power Transformers using a Synthetic Analysis Method Based on Association Rule and Variable Weight Coefficients", IEEE Transactions on Dielectrics and Electrical Insulation, Vol. 20, 2013.

[8] J. R. Daher, S. Chilkaka, A. Younes and K. Shaban, "Association Rule Mining Five Years of Motor Vehicle Crashes", 5th International Conference on Transportation and Traffic Engineering, 2016.

[9] H. Oliff and Y. Liu, "Towards Industry 4.0 Utilizing Data-Mining Techniques: A Case Study on Quality Improvement", The 50th CIRP Conference on Manufacturing Systems, 2017.

[10] N. Liu and L. Ma, "Research of Improved Apriori Algorithm Based on Itemset Array", Sensors and Transducers, Vol. 153, No. 6, pp. 84-91, 2013.

[11] B. Rokaha and D. P. Ghale, "Enhancement of Supermarket Business and Market Plan by Using Hierarchical Clustering and Association Mining Technique", International Conference on Networking and Network Applications, 2018.

[12] R. Ismail, Z. Othman and A. A. Bakar, "Associative Prediction Model and Clustering for Product Forecast Data", 10th International Conference on Intelligent Systems Design and Applications, 2010.

[13] I. H. Witten, E. Frank and M. A. Hall, "Data Mining Practiccal Machine Learning Tools and Techniques", Second Edition, Elsevier, 2005.

[14] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald and D. Scuse, "WEKA Manual", Version 3-6-10, University of Waikato, 2013.

[15] K. Mani and R. Akila, "Enhancing the Performance in

Generating Association Rules using Singleton Apriori", International Journal Of Information Technology and Computer Science, Vol. 9, No. 1, pp. 58-64, 2017.

Authors' Profiles

Ashraf Mohamed ABUSIDA: A native of Libya, received his Bachelor degree in Computer Science in 1996 from the Faculty of Science, University of Tripoli, Tripoli, Libya. From 1998 to 2018, he worked as a software developer for General Electricity Company of Libya (GECOL). Until he became head of the company's developers team. In 2010, he started a Master study of Computer Science in Libyan Academy, Tripoli, Libya. In 2019, he enrolled in a PhD program in Computer Engineering, at the Department of Computer Engineering, University of Kastamonu, Turkey.

Yasemin GÜLTEPE: She received her B.Sc. in Computer Science Department from Çanakkale Onsekiz Mart University, Çanakkale/Turkey in 2000. She received her M.Sc. degree from Natural and Applied Sciences, Çanakkale Onsekiz Mart University, Çanakkale/Turkey and Ph.D. degree from Natural and Applied Sciences, Ege University, Izmir/Turkey. Her research interests include semantic web, linked data, data management, algorithms and medical informatics.

How to cite this paper: Ashraf Mohammed Abusida, Yasemin Gültepe, "An Association Prediction Model: GECOL as a Case Study", International Journal of Information Technology and Computer Science(IJITCS), Vol.11, No.10, pp.34-39, 2019. DOI: 10.5815/ijitcs.2019.10.05