Analysis of Pre-Weighted and Post-Weighted Association Rule Mining

Ayse Betul Cengiz
Graduate School of Natural and Applied Sciences
Dokuz Eylul University
Izmir, Turkey
ayse.simsek@ceng.deu.edu.tr

Kokten Ulas Birant
Department of Computer Engineering
Dokuz Eylul University
Izmir, Turkey
ulas@cs.deu.edu.tr

Derya Birant
Department of Computer Engineering
Dokuz Eylul University
Izmir, Turkey
derya@cs.deu.edu.tr

***Abstract*—Association Rule Mining (ARM) is an important task of data mining that finds interesting patterns and relationships among items in the huge amount of data stored. While widely used in many different fields, the classical ARM assumes that all items have the same significance. Weighted Association Rule Mining (WARM) overcomes this problem by assigning weights to items with the goal of reflecting their importance to the mining results. In this case, the main question is whether weighting process should be applied before ARM or after it. To answer this question, this paper analyzes and compares two alternative approaches: Pre-Weighted Association Rule Mining (PreWARM) and Post-Weighted Association Rule Mining (PostWARM). It is shown by experimental studies that PostWARM produces more compact rules with higher information content and PreWARM finds more meaningful rules than standard rule generation methods.**

Keywords—association rule mining, weighted association rules, data mining

I. INTRODUCTION

Association rule mining (ARM) is a technique that identifies frequent patterns, correlations, or causal structures from large datasets. Association analysis is one of main problems in the field of data mining and plays an important role in a wide variety of application domains such as market basket analysis, recommendation systems, web usage mining, intrusion detection, health and bioinformatics.

However, the traditional ARM has several drawbacks: (i) it assumes that items in data have the same significance without taking account of their importance within a transaction or within the whole item domain, and (ii) a huge number of rules are generated, most of which are of little or no interested for decision making. Clearly these are undesirable, and a different method is required to solve this problem. To address these problems, *Weighted Association Rule Mining* (WARM) was proposed as a solution, where each item is assigned a weight with respect to some user defined criteria. Assigning higher weights to items enables them to appear on more meaningful rules to be discovered.

In this study, two alternative approaches have been analyzed comparatively: *Pre-Weighted Association Rule Mining* (PreWARM) and *Post-Weighted Association Rule Mining* (PostWARM). The first is to obtain rules based on a weighted support threshold from weighted items and transactions by using a weighted ARM algorithm such as the Weigthed Eclat (WEclat) algorithm. The other one is to obtain the rules based on a standard minimum support threshold from non-weighted items and transactions, after that, the weighted support values of these rules are calculated according the item weights and the elimination process is made according to the minimum weighted support threshold. In the experimental studies, two publicly available datasets were used to evaluate the approaches.

The novelty and main contributions of this paper are three-fold: (i) it defines two novel terms PreWARM and PostWARM, (ii) it is the first study that compares pre-weighting and post-weighting approaches for ARM, and (iii) it presents experimental studies conducted on two different real-world datasets to demonstrate that WARM produces more compact rules with higher information content than standard rule generation methods.

The remainder of this paper is organized as follows: In Section 2, related works on the subject are given. In Section 3, background information about ARM and WARM are explained. In addition, this section also describes two alternative approaches (PreWARM and PostWARM) in detail. Section 4 explains the datasets used in the experimental studies and presents the obtained results with discussions. Finally, Section 5 provides some concluding remarks and possible future works.

II. LITERATURE REVIEW

The concept of association rule was first introduced by Agrawal et al. [1]. They defined the support and confidence measurements and proposed association rule mining technique for the discovery of frequent patterns. In 1994, the Apriori algorithm, which is one of the most popular and widely used ARM algorithms, was purposed for mining association rules. The name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties in the previous pass to generate the candidate itemsets. In the following years, many different ARM algorithms (i.e. FP-Growth, Eclat) have been proposed to discover the rules faster. These algorithms produce the same results with the same minimum support and minimum confidence values given by the user.

However, the standard ARM algorithms ignore the differences among items, so they assume that all items have the same significance without taking account of their weights. However, this is not always the case, since different items have generally different weights in real-world datasets. For example, in web mining, dwelling time indicates the page importance, so it can be used to assign weights to the different web pages. Likewise, in a market basket application, item profit is important, so this information can be used as a measure of item weight. For instance, the association rule TV→Laptop may be more important than the rule Speaker→Headphone even though the former holds a lower support. This is because those items in the first rule usually come with more profit per unit sale, but the traditional ARM algorithms ignore this significance. As a result of this drawback, the concept of weighted ARM

978-1-7281-2868-9/19/$31.00 ©2019 IEEE

(WARM) has emerged as an interesting new research direction. Differently from ARM, WARM assigns weights to items or transactions for differentiating them.

In recent years, WARM has been used in different application areas such as medicine [2], [3], bioinformatics [4], sport [5], finance [6], education [7] and manufacturing [8]. In each application of WARM, how weights are given is a matter of debate. Differently from the previous works in the literature, our study investigates the performance of two possible weighting approaches: Pre-Weighted Association Rule Mining (PreWARM) and Post-Weighted Association Rule Mining (PostWARM).

III. Materials and Methods

A. Association Rule Mining

Given a set of m distinct items denoted by $I = \{i_1, i_2, ..., i_m\}$. Let dataset D be a set of transactions, where each transaction T consists of a set of items such that $T \subseteq I$. A set of items (i.e. X or Y) is called an *itemset*. The number of items in an itemset is called the *length* of the itemset. Itemset of some length k is referred to as a *k-itemset*. Given a set of transactions, ARM aims to find rules of the form $X \rightarrow Y$, where X, Y are two sets of items, $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. The meaning of the rule is that if the left hand side (X) occurs, then the right-hand side (Y) is also very likely to occur. The interestingness of the rules is often measured using support and confidence values. The *support* of a rule is defined as the number of records in the dataset that contain both X and Y (Eq. (1)). In other words, support is the fraction of transactions containing the itemset. The *confidence* of a rule is defined as the proportion of records containing Y among those records containing X (Eq. (2)). In other words, confidence is probability of occurrence of Y given X is present.

$$Support(X \rightarrow Y) = \frac{Number\ of\ transactions\ with\ both\ X\ and\ Y}{Total\ number\ of\ transactions} \quad (1)$$

$$Confidence(X \rightarrow Y) = \frac{Number\ of\ transactions\ with\ both\ X\ and\ Y}{Total\ number\ of\ transactions\ with\ X} \quad (2)$$

In order to select interesting rules from the set of all possible rules, user defined thresholds are used. The best-known thresholds are minimum support (*minSP*) and minimum confidence (*minCF*). The task of ARM is to generate all association rules that have support and confidence above these user-specified thresholds.

B. Weighted Association Rule Mining

The classical ARM employs the support measure, which treats every item equally. In contrast, different items have generally different weights in real-world datasets. This led to the emergence of weighted association rule mining (WARM) concept. In WARM, a weight w_i is assigned to each item i, reflecting the relative importance of an item over other items. The weight of an itemset is derived from the weights of the items present in the rule or in the transaction as given in (3):

$$weight(IS_k) = \frac{\sum_{i \in IS_k} w_i}{|IS_k|} \quad (3)$$

where IS refers to an itemset, IS_k is the k^{th} itemset in the list, and $|IS_k|$ denotes to the number of items in the itemset.

C. Pre-Weighted and Post-Weighted Association Rule Mining

In this study, we introduce two novel terms: Pre-Weighted Association Rule Mining (PreWARM) and Post-Weighted Association Rule Mining (PostWARM). While PreWARM applies weighting process before ARM, PostWARM utilizes after it.

Table I presents the main stages of two alternative approaches: PreWARM and PostWARM. They both extract the rules that contain significant items, since they consider the items with higher weights.

TABLE I. The Main Stages of PreWARM and PostWARM Methods

Steps	Post-Weighted ARM	Pre-Weighted ARM
i.	Select minimum support threshold	Assign weight values to items
ii.	Run Eclat algorithm and find the rules with their support values	Calculate transaction weights from the weights of the items present in the transaction (Eq. 3)
iii.	Assign weight values to items	Select minimum weighted support threshold
iv.	For each rule, calculate the rule weight from the weights of the items present in the rule (Eq. 3)	Run "Weighted Eclat" algorithm and find the rules (Eq. 4)
v.	Calculate the weighted support values of the rules (Eq. 5)	
vi.	Eliminate the rules whose weighted support is less than given minimum weighted support threshold	

PreWARM is to obtain rules based on a weighted support threshold from weighted items and transactions by using a weighted ARM algorithm such as the Weighted Eclat (WEclat) algorithm. Firstly, Equation (3) is applied to find the weight of each transaction (*weight(T)*). Then, the transaction weights are used to find frequent itemset. With this concept, support term has also evolved to weighted support. To calculate the weighted support (*WSP*) of an itemset, the weights of the transactions that contain this itemset are summed and divided by the sum of the weights of all transactions as given in Equation (4):

$$WSP(IS_k) = \frac{\sum_{IS_k \subseteq T} weight(T)}{\sum_{T \in D} weight(T)} \quad (4)$$

where D is the dataset, T is a transaction in the dataset D, IS refers to an itemset, and IS_k is the k^{th} itemset in the list. An itemset is frequent if its weighted support is equal or greater than the minimum weighted support threshold given by the user.

PostWARM is initially finds the rules by using a standard ARM algorithm such as Eclat, and then applies a weighting process. Firstly, the association rules are found based on a particular minimum support (*minSP*) threshold from non-weighted items and transactions. After that, for each rule, the weight of the rule is calculated by from the weights of the items present in this rule by using Equation 3. Then, the weighted support (*WSP*) value of the rule is computed by multiplying the rule weight with support value [9] as given in Equation (5):

$$WSP(R_k) = support(R_k) * weight(R_k) \quad (5)$$

where R refers to rule and R_k is the k^{th} rule in the rule list. Finally, the elimination process is made according to the minimum weighted support threshold (*minWSP*) given by the user such that $WSP \geq minWSP$.

D. Examples for Pre-WARM and Post-WARM Methods

1) Example for Post-WARM Approach

Step 1: User determines a minimum support threshold (*minSP*). As shown in Table II, other parameters (i.e. the maximum and minimum number of items) can also be determined.

TABLE II. PARAMETERS USED TO DISCOVER RULES

Parameters	Values
Minimum Support	0.2
Minimum Rule Length	2
Maximum Rule Length	10

Step 2: By running a standard ARM algorithm such as Eclat, association rules with support values equal to or greater than *minSP* are discovered. Table III shows an example rule.

TABLE III. A SAMPLE OF RULE OBTAINED BY THE ECLAT ALGORITHM

Rule	Support	Confidence
{Monsoon Wedding, Terminator 3: Rise of the Machines} => {Solaris}	0.2309985	0.8072917

Step 3: As shown in Table IV, a weight is assigned to each item, reflecting the relative importance of an item over other items. Weights can be determined according to the problem such as item prices, ratings or frequencies.

TABLE IV. SAMPLE ITEMS AND THEIR WEIGHTS

Symbol	Items (Movie)	Weight
A	Solaris	4.138158
B	Monsoon Wedding	3.706204
C	Terminator 3: Rise of the Machines	4.256173

Step 4: To calculate the weight of the rule R, find the average of the weights of the items present in the rule (Eq. 3).

$$\text{weight(R)} = \frac{\text{weight(A)} + \text{weight(B)} + \text{weight(C)}}{\text{number of items in R}}$$

$$= \frac{4.138158 + 3.706204 + 4.256173}{3} = 4.033511$$

Step 5: As shown in Table V, The weighted support value of the rule is calculated (Eq. 5).

$$WSP(R) = \text{support(R)} * \text{weight(R)}$$

$$= 0.2309985 * 4.033511 = 0.931734$$

TABLE V. CALCULATED RULE WEIGHT AND WEIGHTED SUPPORT VALUES

Rule	Support	Rule Weight	Weighted Support
{Monsoon Wedding, Terminator 3: Rise of the Machines} => {Solaris}	0.2309985	4.033511	0.931734

Step 6: The rules whose weighted support (*WSP*) is less than given minimum weighted support (*minWSP*) threshold are eliminated.

2) Example for PreWARM Approach

PreWARM approach uses a weighted ARM algorithm such as Weighted Eclat (WEclat). The WEclat algorithm generates rules based on transaction weights. Table VI shows a sample rule discovered by the WEclat algorithm.

TABLE VI. A SAMPLE RULE DISCOVERED BY THE WECLAT ALGORITHM

Rule	Support	Confidence
{To Kill a Mockingbird}=>{Silent Hill}	0.2116244	0.8068182

IV. EXPERIMENTAL STUDIES

To be able to compare PreWARM and PostWARM approaches, we implemented them in the R language.

A. Dataset Description

In the experimental studies, two different datasets were used to evaluate the alternative approaches: Movies Dataset and Online Retail Dataset.

i. Movies Dataset: This dataset is provided by Kaggle [10] and consists of several files. Among these files, only the files given in Table VII were used. The first file contains information on 45,000 movies featured in the Full MovieLens dataset. The second file includes the subset of 100,000 ratings from 700 users on 9,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website. The dataset consists of movies released on or before July 2017.

TABLE VII. MAIN FEATURES IN THE MOVIES DATASET

File Name	Features
Movies metadata	adult, budget, genres, belongs to collection, homepage, id, imdb_id, original_title, spoken languages, original_language, overview, popularity, poster_path, production_companies, production countries, release_date, revenue, status, runtime, tagline, title, video, vote_count, vote_avg
Ratings small	userId, movieId, rating, timestamp

ii. Online Retail Dataset: This dataset is provided by UCI [11]. It contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The dataset consists of 541,909 instances and 8 attributes as given in Table VIII.

TABLE VIII. CHARACTERISTICS OF ONLINE RETAIL DATASET

Features	Description	Type
InvoiceNo	Invoice number, uniquely assigned to each transaction	Nominal
StockCode	Product (item) code, uniquely assigned to each distinct product	Nominal
Description	Product (item) name	Nominal
Quantity	The quantities of each product (item) per transaction	Numeric
InvoiceDate	The day and time when each transaction was generated	Numeric
UnitPrice	Product price per unit in sterling	Numeric
CustomerID	Customer number, uniquely assigned to each customer	Nominal
Country	The name of the country where each customer resides	Nominal

B. Data Preparation

1) Data Preparation for the Movies Dataset

Firstly, the empty rows in the dataset were deleted. Items (movies) were weighted in two different ways.

- Average rating of a movie (Average): The average rate for each movie is calculated and is used as weight.
- Number of views of a movie (Count): It is calculated how many people watch each movie and this value is used as weight.

2) Data Preparation for Online Retail Dataset

- Empty cells were ignored.
- Canceled invoices, whose code starts with letter 'c', were removed from the dataset.
- The invoice amounts were used as the weights.

C. Experimental Results

As shown in Table IX, when the support and confidence values were kept constant for the same dataset, there was a difference in item set and rule numbers obtained from different weighted datasets. However, it is possible to see increase of rule numbers in the pre-Weighted method, although there was no increase of rule numbers in post-Weighted method.

TABLE IX. SAMPLE EXPERIMENTAL RESULTS OBTAINED FROM THE MOVIES DATASET

Weights	Algorithm	Min. Support	Min. Confidence	#Itemset	#Rules
-	ARM	0.2	0.8	69	9
Average	Post-WARM	0.2	0.8	45	6
Average	Pre-WARM	0.2	0.8	75	10
Count	Post-WARM	0.2	0.8	58	8
Count	Pre-WARM	0.2	0.8	87	14

Fig. 1 and Fig. 2 show the comparative results obtained by ARM, PreWARM and PostWARM algorithms from the Movies dataset in various minimum support values for average weighting and count weighting strategies respectively. It is observed that, when the minimum support value increases, the number of rules decreases similarly in all three methods. Since the PostWARM method weights and eliminates the rules after ARM process, the number of rules is less than the other methods in all minimum support values.

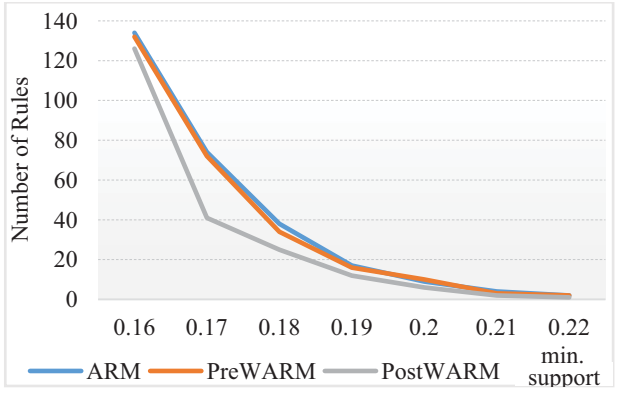

Fig. 1. Changing the number of rules according to minimum support for average weighing.

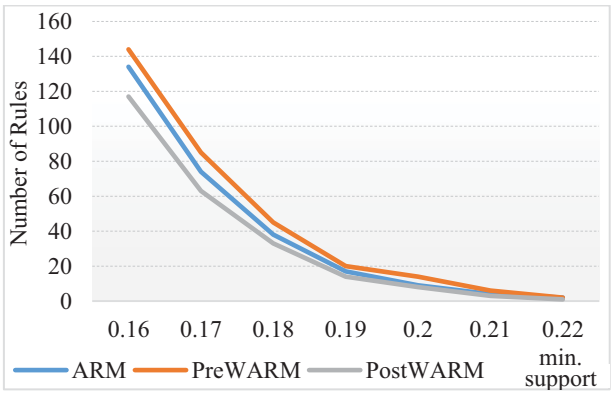

Fig. 2. Changing the number of rules according to minimum support for count weighting.

Fig. 3 and Fig. 4 show the number of rules obtained for two different weighting strategies (average weighting and count weighting) by PreWARM and PostWARM algorithms respectively. The results clearly indicate that, weighting strategy has a direct effect on the number of rules. It is observed that the number of frequent rules produced by count based weighting strategy is higher than the rules obtained by average based weighting mechanism.

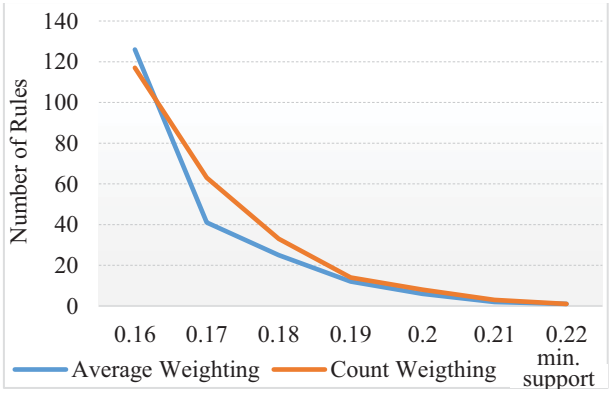

Fig. 3. The number of rules discovered by PostWARM method for count based weighing and average based weighing strategies.

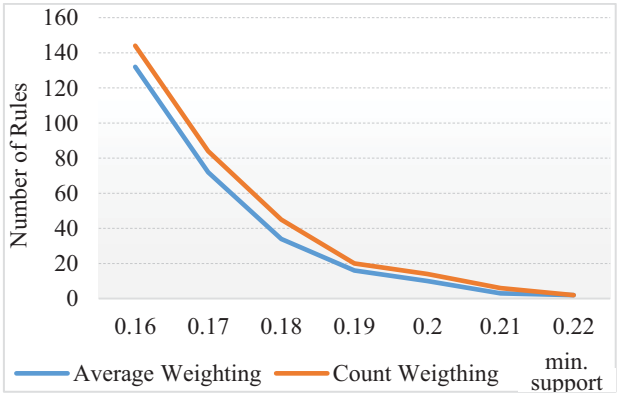

Fig. 4. The number of rules discovered by PreWARM method for count based weighing and average based weighing strategies.

Table X shows the changes in the number of frequent itemsets and rules discovered by varying the minimum support. The WEclat algorithm produces more reasonable

number of frequent itemsets and rules. Thus, it is possible to find the association rules in a more manageable size.

TABLE X. SAMPLE EXPERIMENTAL RESULTS OBTAINED FROM THE ONLINE RETAIL DATASET

Weights	Algorithm	Min. Support	Min. Confidence	#Itemset	#Rules
-	Eclat	0.008	0.6	731	385
Amount	WEclat	0.008	0.6	39	24
-	Eclat	0.01	0.6	353	138
Amount	WEclat	0.01	0.6	19	16
-	Eclat	0.012	0.6	212	78
Amount	WEclat	0.012	0.6	8	10
-	Eclat	0.015	0.6	87	31
Amount	WEclat	0.015	0.6	3	4

Fig. 5 shows the graph that compares Eclat and WEclat algorithms for the Online Retail dataset with varying minimum support and 0.6 minimum confidence thresholds. It can be clearly observed that the Eclat algorithm always discovers more frequent rules than the WEclat algorithm. This is due to the effect of item weights. This effect is significant since too many frequent rules are found with small *minSP* and *minCF* thresholds. For example, when minSP is 0.01, Eclat finds 138 rules while WEclat discovers 16 frequent rules.

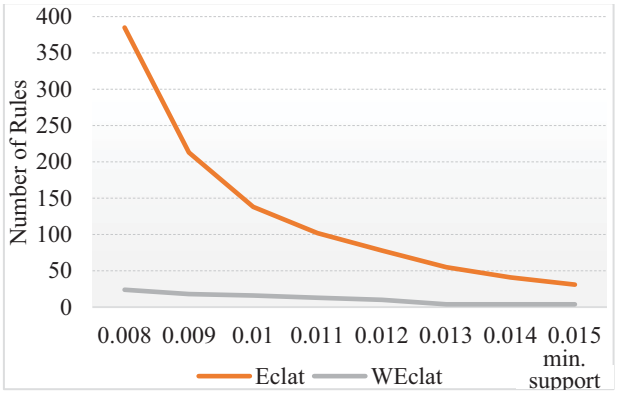

Fig. 5. Comparison of Eclat and WEclat algorithms.

V. CONCLUSION AND FUTURE WORKS

Since the traditional ARM algorithm ignores the importance of items, the WARM approach has come to the fore in recent years. More meaningful rules can be formed with the help of weight values given to items. However, WARM approach raises several questions that should be answered: when and how the items should be weighted. In this study, two alternative approaches called PreWARM and PostWARM were tested to answer these questions. While PreWARM approach considers the item or transaction weights from the first stage, PostWARM method works like a rule filter.

In the experimental studies, two different datasets were used to evaluate the alternative approaches: Movies Dataset and Online Retail Dataset. It is shown by experimental studies that PostWARM produces more compact rules with higher information content. It is also possible to say that PreWARM finds more meaningful rules with the help of weight values since the weights indicate the significance of the items.

In the future, several extensions can be made to improve the study. In the PostWARM approach, rule weights are multiplied by their support values, so the result can reach to high ranges. To prevent this, it may be necessary to normalize the values in order to keep the weight values within a certain range. Another point is the time interval at which transactions are made. The development of time-based approaches in the future may allow time-based weighting.

REFERENCES

[1] R. Agrawal, T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases," *Proceedings of the ACM SIGMOD international conference on management of data*, *25-28 May 1993*, Washington, D.C. / USA, ACM SIGMOD Record, vol. 22 (2), pp. 207-216, 1993.

[2] X. Zhu, Y. Liu, Q. Li, Y. Zhang, C. Wen, "Mining effective patterns of Chinese medicinal formulae using top-k weighted association rules for the internet of medical things," *IEEE Access*, vol. 6, pp. 57840-57855, 2018.

[3] K. S. Lakshmi, G. Vadivu, "A novel approach for disease comorbidity prediction using weighted association rule mining," *Journal of Ambient Intelligence and Humanized Computing*, (in press).

[4] M. Indhumathy, A. R. Nabhan, S. Arumugam, "A weighted association rule mining method for predicting HCV-human protein interactions," *Current Bioinformatics*, vol. 13 (1), pp. 73-84, 2016.

[5] D. Saraswat, V. Dev, P. Singh, "Analyzing the performance of the indian cricket team using weighted association rule mining," *Int. Conf. Comput. Power Commun. Technologies*, *28-29 Sept. 2018*, Uttar Pradesh / India, pp. 161-164, 2018.

[6] C. H. Cheng, C. H. Chen, "Fuzzy time series model based on weighted association rule for financial market forecasting," *Expert Systems*, vol. 35 (4), pp. 1-15, 2018.

[7] H. C. Chen, J. L. Zhang, Y. Q. Deng, "Application of mixed-weighted-association-rules-based data mining technology in college examination grades analysis," *International Journal of Digital Content Technology and its Applications*, vol. 6 (10), pp. 336-344, 2012.

[8] Y. He, C. Zhu, Z. He, C. Gu, and J. Cui, "Big data oriented root cause identification approach based on axiomatic domain mapping and weighted association rule mining for product infant failure," *Computers & Industrial Engineering*, vol. 109, pp. 253-265, 2017.

[9] F. Tao, F. Murtagh, M. Farid, "Weighted Association Rule Mining using Weighted Support and Significance Framework," *Proc. of the 9th ACM SIGKDD int. conf. on knowledge discovery and data mining, 24-27 August 2003*, Washington, D.C / USA, pp. 661-666.

[10] R. Banik, "The Movies Dataset." [Online]. Available: https://www.kaggle.com/rounakbanik/the-movies-dataset.

[11] D. D. Chen, "Online Retail Dataset." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/online+retail