VOL.5NO.1 March 2016 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

A Literature Survey on Association Rule Mining Algorithms

Pınar Yazgan[a], Ali Osman Kuşakcı[b]

[a,b]Istanbul Commerce University, Graduate School of Natural and Applied Sciences, Küçükyalı E5 KavşağıInönü Cad. No: 4, Maltepe 34840, Istanbul, Turkey
pinar.yazgan@windowslive.com, aokusakci@ticaret.edu.tr

Article Info
Article history:
Article received on 8 Jan. 2016
Received in revised form
18Mar.2016

Keywords:
Information technology, market basket analysis, association rule mining, data mining.

Abstract
With the development of database technology, the need for data mining arises. As a result, Association Rule Mining(ARM) has become a very hot topic in data mining. This paper presents definition and application areas of association rules. Furthermore, a comprehensive literature review on the existing algorithms of ARM is conducted with a special focus on the performance and application areas of the algorithms. These algorithms are in general classified into three main classes: (1) based on frequent itemset, (2) based on sequential pattern, and (3) based on structured pattern. The algorithms are developed to improve the accuracy and decrease the complexity, and execution time. However, it is hard to say that they do always succeed to optimize all these aspects simultaneously. Hence, there is still some space to develop more efficient algorithms for different data structures.

1. INTRODUCTION

Markets, companies and other organizations with the help of rapidly evolving information technology (IT) in recent years, have the opportunity to store huge data. Data series are constructed by keeping record of all operations within the organizations [1]. Evaluation of day by day growing datasets and extracting useful information is of paramount importance. Today, IT allows acquiring and keeping a huge amount of data. However, the bulk data comes along with a difficult task: to analyze this huge data and obtain the correct information. Obviously, making an analysis by observing the tables with thousands of lines and columns and making useful conclusions is practically impossible. This necessitates employing computers with vast capacity running fast algorithms. Finding patterns, trends and abnormalities in datasets and summarizing them as simple models is one of the most important issues in the information age[2].

Data mining is discovery of relations and rules which is significant, potentially useful and making predictions about future through large amount of available data using computer programs [3].

One of the data mining application areas, in increasingly widespread use in many sectors, is Market Basket Analysis (MBA) in which the relationship and the rules are obtained taking advantage of the customer, product and sales information. MBA, obtaining of products' sales relationship with another product and finding out the data mining association rules, increases the profits of companies.

Association rules provide rules for the generation of future predictions discovering objects acting together in a sales transaction data. To obtain these rules, since the beginning of the 90s, many algorithms have been developed. Each of them has different working principles and advantages relative to each other under different circumstances. Implementation of merge and pruning methods, scanning database and presence of association relationship between objects represent general procedures of the algorithms.

2. ASSOCIATION RULE MINING

Association rules have been an attractive research topic of data mining and attract more attention of researchers with the help of gradually increasing computational power achieved in the recent decade. Simply said, they are used for discovering frequent patterns, associations, correlations between data.

Association rule mining(ARM) may be helpful in the task of decision making by finding relationships among the attributes of a database. For example: new useful rules are

gained through the sales transaction database which consist of purchasing behavior of customers can be found with ARM [4].

Association rules have been used in many areas, such as:

Market Basket Analysis: MBA is one of the most typical application areas of ARM. When a customer buys any product, what other products s/he puts in the basket with some probability may be determined by applying association rules.

This new knowledge can be exploited by store manager to organize the shelves accordingly. Thus, customers can reach these products more easily. This, in turn, results with an increase in sales rates.

Medical diagnosis: Association rules can be used for helping doctors to treat patients. Serban et al. propose a technique based on relational association rules aiming to determine the probability ofa certain illness [5].

Protein Sequences Proteins are sequences consist of 20 different amino acids. Each protein is made up of a unique 3-dimensional structure with amino-acid sequence. Gupta et al. find associations between different amino acids in a protein [6].

Census Data: Censuses make a wide range of general statistical information on society. Public services such as health, education, transport and public business can take advantage of the information related to population and economic census for planning [7].

Malerba et al. suggest a method for finding spatial association rules involving relations among objects in census data [7].

Customer Relationship Management (CRM) of Credit Card Business: CRM helps to develop the relationship between credit card customers and the bank with identifying the preference of various products, services and customer groups according to their choices. Chen et al. classified customers into groups to determine high-profit or so called gold customers [8].

2.1 Basic Measures

The two important basic measures of association rules are support and confidence.

Support defines how often a rule is applicable to a given data set. The rule X ⇒Y holds with support s if s% of transactions in D consists of X ∪Y.

$$S = \frac{\sigma(X \cup Y)}{\#\ of\ trans.} \quad (1)$$

In table 1, a simple case with five instances is used to demonstrate how the support of a certain X=>Y rule can be calculated.

Confidence identifies how frequently items in Y found in transactions that consist of X. The rule X⇒Y holds with confidence probability of c given that c% of the transactions in D that consist of X also contain Y.

$$C = \frac{\sigma(X \cup Y)}{\sigma(X)} \quad (2)$$

In table 2, a simple case with five instances is used to demonstrate how the confidence of a certain X=>Y rule can be calculated.

TABLE 1: EXAMPLE OF SUPPORT MEASURE

TID	Items	Given X⇒Y Confidence{X⇒Y}=Occurrence {Y} / Occurrence {X}
1	Bread, Butter, Peanut	
2	Bread,Butter, Milk	C{Bread⇒Butter}=2/3=%66
3	Butter,Peanut	C{Butter⇒Peanut}=3/4=%75
4	Bread,Peanut	C{Bread,Butter⇒Peanut}=1/2=%50
5	Butter,Peanut,Milk	

TABLE 2: EXAMPLE OF CONFIDENCE MEASURE

TID	Items	Support{X⇒Y}=Occurrence /Total Support
1	Bread, Butter, Peanut	
2	Bread,Butter, Milk	Total Support=5 Support{Bread,Butter}=2/5=%40 Support{Butter,Peanut}=3/5=%60 Support{Bread,Butter,Peanut}=1/5=%20
3	Butter,Peanut	
4	Bread,Peanut	
5	Butter,Peanut,Milk	

3. LITERATURE OVERVIEW

This section presents a literature review on different techniques for ARM with a special attention on MBA applications.

ARM algorithms can be classified into three main classes: (1) Frequent itemset mining, (2) Sequential pattern mining, (3) Structured pattern mining. Figure 1 shows various association rule mining algorithms with sub branches developed since the first introduction of ARM algorithms.

3.1. Frequent Itemset Mining

Today, frequent itemset mining is one of the most important tools employed on transactional database and has an important role in many data mining tasks to discover patterns such as classifiers, correlations, clusters, association rules, sequences. It aims to optimize the process of finding patterns in a dataset.

A frequent itemset is a pattern that is observed more frequently than a certain threshold, minimum support, in the database. Some strategies used to generate frequent itemsets are:

- Reduce the number of candidates (M) by using pruning techniques to decrease M.
- Decrease the number of transactions (N) with using Direct Hashing and Pruning (DHP) and vertical-based mining algorithms.
- Decrease the number of comparisons (NM) with using efficient data structures for storing transactions or candidates.

Possible applications of algorithms based on frequent item sets approach are:

- Develop arrangement of products on a catalog's pages,in shelves,
- Product bundling, support cross-selling applications,

- Technical dependence analysis, fraud detection[9].

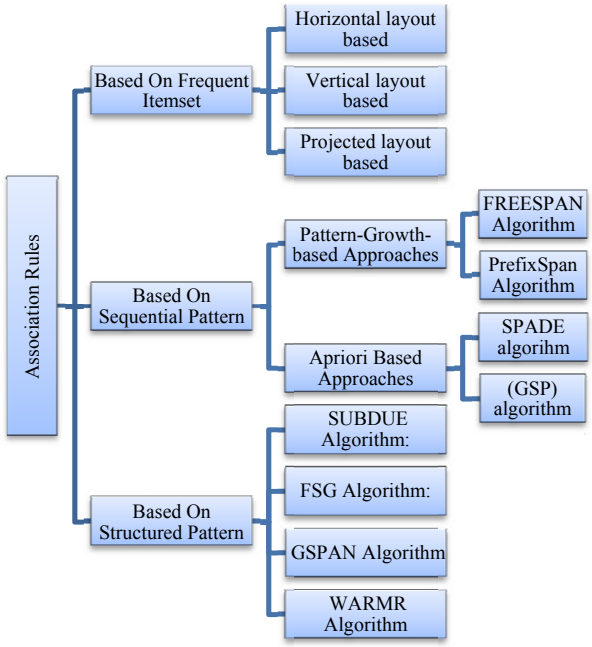

Figure 1.Taxonomy of Association Rule Mining Algorithms

TABLE 3: ILLUSTRATIVE TRANSACTION DATABASE

Transaction Database
1:{milk, diaper, bread}
2:{beer, coke, diaper}
3:{milk, coke, bread}
4:{ milk, coke, diaper, bread}
5:{milk, bread}
6:{milk, coke, diaper}
7:{beer, coke}
8:{milk, coke, diaper, bread}
9:{beer, coke, bread}
10:{milk, diaper, bread}

TABLE 4 FREQUENT ITEMSETS

0 item	1 item	2 items	3 items
0: 10	{milk}:7 {beer}:3 {coke}:7 {diaper}:6 {bread}:7	{milk,coke}: 4 {milk,diaper}: 5 {milk,bread}: 6 {beer,coke}: 3 {coke,diaper}: 4 {coke,bread}: 4 {diaper,bread}: 4	{milk,coke,diaper}: 3 {milk,coke,bread}: 3 {milk,diaper,bread}: 4

We can illustrate the main steps of finding Frequent Itemsets on a transaction database through the following example. Assume we have a list of 10 transaction and frequent itemsets given in Table 3 and Table 4.

The minimum support is set as s_{min} = 3 orσ_{min} = 0.3 = 30%. According to Table 4, following can be concluded;

• There are 2^5 = 32 possible item sets over this 5 different products given in these 10 transactions. B={milk, beer, coke, diaper, bread}

• For the specified minimum support level, there are 16 frequent item sets but only 10 transactions.

Many algorithms for mining frequent itemsets have been introduced over the years like horizontal layout based algorithms, vertical layout based algorithms and projected layout based algorithm. These algorithms are shown in Table 7.

3.1.1 Algorithms Based On Horizontal Layout Database

In this type of database, each row records a transaction with a transaction identifier (TID), followed by a set of items purchased during that transaction. An example of horizontal layout dataset is shown in Table 5 [10].

TABLE 5: EXAMPLE OF HORIZONTAL LAYOUT DATABASE

TID	**ITEMS**
T1	I1,I2,I4,I6,I7
T2	I1,I2,I4,I8
T3	I1,I3,I4,I6,I9
T4	I2,I4,I11
T5	I6,I8
T6	I1,I3,I4

Many techniques have been proposed to mine frequent patterns from horizontal data format such as Apriori Algorithm, Direct hashing and pruning (DHP), Partitioning algorithm, Sampling algorithm, Continuous Association Rule Mining Algorithm (CARMA), Dynamic Itemset Counting (DIC).

Apriori Algorithm

Apriori algorithm has been developed by Agarwal and Srikant in 1994.It is one of the most famous of all ARM algorithms. Apriori is designed to operate on databases containing transactions and can be used to generate all frequent itemsets.

Along with Apriori, AprioriTid and AprioriHybrid algorithms have been suggested by Agrawal and Ramakrishnan in 1994. AprioriTid algorithm is executed equivalently well as Apriori in small problems, but when implemented to big-scale problems performance of the algorithm decreases. On the other hand, AprioriHybrid performs better than Apriori in almost all cases[11].

Many improvements have been made on the Apriori algorithm in order to increase its efficiency and effectiveness [12]. Apriori algorithm is not sufficient for redundancy and candidate set generation. However, it forms the basis for many algorithms.

Ji et al. presented a development on the existing Apriori algorithm. Unwanted candidate set generations are the disadvantage of existing Apriorialgorithm. Modifying the pruning technique decreased the candidate set generation process in improved Apriori algorithm. A separate set is developed with less frequent items. This improvement has notable advantages over the standardApriori Algorithm [13].

Huang has changed Apriori Algorithm without large candidate set generation [14].Wu et al proposed another Improved Apriori Algorithm (IAA) to decrease the number of scans on the data and redundant operations while frequent itemsets and association rules are produced. This algorithm has a new count-based method for pruning process. In this algorithm, all frequent itemsets are found in given data sets using genetic algorithm [15].

Along with the rapid increase inlog-data, there was a need for handling this type of data. Shao et al. suggested 3D-Apriori algorithm. 3D-Apriori algorithm has main features as attribute data discretion and spatial predicate extraction for generation of association rules. 3D-Apriori interprets the logging data and enhances efficiency of association rules behind the logging data transformation [16].

Sharmaand Sufyan developed a probabilistic analysis of Apriori algorithm to discover association rules and frequent itemsets in a transactional database. It contains a single database scan and limit unsuccessful candidate sets. The concept of recursive medians is used in the algorithm as an Inverted V-Median Search Tree (IVMST). The recursive medians compute the dispersion of each itemsets in the transaction list and the maximum number of common transactions for any two itemsets. Using the above mentioned procedures, they presented a time efficient algorithm to discover frequent itemsets [17].

Wang et al. improved the efficiency of data mining in large transaction database by applying Fast-Apriori algorithm. According to the authors, data mining engine could be derived using an integration of various mining algorithms, cluster analysis, regression analysis, classification and other techniques. An engine, obtains queries from users, search the memory to show suitable results to the user. They applied a fast approach into existing Apriori algorithm for getting quick responses. This fast algorithm has better performance than Apriori algorithm [18].

Zeng et al. concentrated on time and space complexity of Apriori algorithm and optimize the complexities. The Hash Mapping Table (HMT) and Hash Tree methodologies were used to optimize time and space complexity. HMT and Hash Tree store transactions and can locate the itemsets easily [19].The authors claim that data collection and evaluation processes are comparatively faster than traditional Apriori algorithm.

Xiaohui proposed a new kind of ARM algorithm and presented an improved Apriori algorithm. The improved algorithm can decrease the input-output operation of mining process and reduce times of database searching, saving storage space required during application of the algorithm [20]. This method is more efficient than the traditional algorithms in mining association rule.

Enhanced scaling Apriori was suggested by Prakash and Parvathi in 2010. This method is an improved Apriori algorithm to limit the number of candidate sets while generating association rules and overall execution time [21].

In 2008,Kamrulet al. presented a novel algorithm, named as Reverse Apriori Frequent Pattern Mining. This algorithm works efficiently and produces large frequent itemsets and reduces the number of items until it takes the largest frequent itemsets [22].

Direct Hashing and Pruning (DHP) Algorithm

DHP algorithm was proposed by Park et al. in 1995todecrease the number of candidates in the early passes and the size of database. DHP employs a hashing technique aiming to restrict the number of candidate itemsets, efficiently generate large itemsets and reduce the transaction database size [23].

Another hash-based approach for mining frequent itemsets was developed by Wang and Chen in 2009. The information of all itemsets is fitted into a structure by using fixed hash-based technique. This method summarizes the data information by using a hash table for predicting the number of the non-frequent itemsets and speeds up the process [24].

Partitioning Algorithm

Partitioning algorithm was proposed by Savasere et al. which is based on idea of partitioning of database in *n* parts to find the frequent elements so that memory problems for large databases can be solved since database is divided into several parts [25].

This algorithm decreases database scan to generate frequent itemsets but time for computing the frequency of candidate generated in each partitions increases. On the other hand, it reduces the I/O overhead and CPU overhead for most cases significantly.

Dynamic Itemset Counting Algorithm (DIC)

DIC algorithm is designed by Brinet al. for database partitions into intervals of a fixed size to reduce the number of transitions through the database.

This algorithm aims to find large itemsets which uses fewer passes over the data than traditional algorithms. Also, DIC algorithm presents a new method of generating implication rules. These rules are standardized based on both the antecedent and the consequent [26].

Sampling Algorithm

Sampling algorithm was presented by Toivonen in 1996. In this algorithm, a sample of itemsets R is taken from the database instead of whole database D.

This algorithm reduces the database activity for finding association rules as it requires only a subsample of the database scanned [27].This algorithm is suitable for any kind of databases, although it sometimes cannot give accurate results.

Continuous Association Rule Mining Algorithm(CARMA)

CARMA was proposed in 1999 by Hidber to compute large itemsets online. This algorithm uses novel method to limit the interval size to 1.The user can change some design parameters, such as minimum confidence, minimum support and the parameters during the first scan of the transaction sequence [28]. CARMA outperforms Apriori and DIC on low support thresholds.

Split and Merge Algorithm (SAM)

SAM algorithm was introduced by Borgelt et al.in 2009. It finds frequent item sets with a split and merge technique where the data is represented as an array of transactions. The traversal order for the prefix tree and the horizontal representation form of the transaction database can be combined. In each step, two subproblems formed with a split step and a merge step in two conditional databases [29].

PRICES Algorithm

This algorithm was created by Wang and Tjortjisin 2004 which firstly recognizes all large itemsets and creates association rules. It reduces large itemset generation time by scanning database and logical operations [30]. It is an efficient method and, in some cases, ten times as fast as Apriori algorithm.

3.1.2.Algorithms Based On Vertical Layout Database

In vertical layout data set, each column corresponds to an item, followed by a TID list, which is the list of rows that the item appears. An example of vertical layout database set is given in Table 6 [10].

TABLE 6: EXAMPLE OF VERTICAL LAYOUT DATABASE

ITEM	TID_List
I1	T1, T4, T5
I2	T3,T5
I3	T1, T2, T4
I4	T3, T4, T5
I5	T1, T2, T3

Equivalence CLASS Transformation Algorithm (ECLAT)

ECLAT algorithm was created by Zaki in 2000 for discovering frequent itemsets from a transaction database. It uses vertical layout.

Each item utilizes intersection based method to calculate the support. Support of an itemset can be calculated by intersecting of any two subsets. Confidence is not calculated in this algorithm [31]. The algorithm finds the elements using depth first search and scans the database only once.

3.1.3. Algorithms for Mining from Projected Layout Based Database

This kind of database uses divide and conquer strategy to mine the useful knowledge. It counts the support more efficiently than based on Apriori algorithms. The projected layout consists of record id separated by column. Tree Projection algorithms may work based on two kinds of ordering: breadth-first and depth-first [32].

FP_Growth Algorithm

FP-growth method has been devised for mining of the complete set of frequent itemsets without candidate generation by Han et al. in 2000.

FP-growth method is an efficient tool to mine long and short frequent patterns [33].We may state several benefits over the other methods:

- Creating a highly compact FP-tree which is smaller than the original database,
- Implementing a pattern growth method to avoid costly candidate generation,
- Saving the costly database scans in the subsequent mining processes,
- And working in a divide-and-conquer way and decreasing the size of the subsequent conditional pattern bases and conditional FP-trees.

Many alternatives and extensions were implemented to the FP-Growth approach: Depth-first frequent itemset generating algorithm[35], H-Mine algorithm[34]; exploring top-down and bottom-up traversal of such trees in pattern-growth mining; and prefix-tree-structure for efficient pattern growth mining.

H_Mine Algorithm

H-Mine was developed by Pei et al. in 2007 which was created using in-memory pointers. H-mine uses an H-struct new data structure for mining[34].In large databases, it firstly makes a partitioning of the database and mines the partitions in main memory using H-struct. It benefits of this data structure and dynamically adjusts links in the mining process and runs very quickly in memory-based settings. H-mine has demonstrated a good performance for various kinds of data. However, its execution time is larger than other algorithms because of the partitioning process [34].

3.2. Sequential Pattern Mining

Sequential pattern mining discovers frequent subsequences as patterns in a sequence database. Ordered elements or events are found in a sequence database. For example: <a(bc)dc> is a *subsequence* of <a(abc)(ac)d(cf)> Table 8 shows a sequence database.

TABLE 8: A SEQUENCE DATABASE

SID	Sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

There are several applications of sequential pattern mining:

- Customer shopping sequences: A customer can make several next purchases, e.g., buying a PC and some Software and Antivirus tools, followed by buying a memory card, and finally buying a printer and some office papers.
- Medical treatments, natural disasters.
- Science and engineering processes.
- Telephone calling patterns, Weblog click streams.
- DNA sequences and gene structures.

Sequential pattern mining can be classified into two major groups: (1)Apriori-based Approaches, and (2) Pattern-Growth-based Approaches.

3.2.1.Apriori Based Approaches(The candidate generation-and-test approach)

The candidate generation-and-test approach is an extension of the Apriori-based frequent pattern mining algorithm to sequential pattern analysis [40].

TABLE 7: FREQUENT ITEMSET MINING

Horizontal Layout Based			
Study	**Authors / Year**	**Method**	**Advantages**
Fast Algorithms for Mining Association Rules [11].	Agrawal, Ramakrishnan, 1994	Combining the best features of Apriori and AprioriTid, AprioriHybrid algorithm.	AprioriHybrid performs better than Apriori in almost all cases.
An Effective Hash based Algorithm for Mining Association Rules [23].	Park et al., 1995	DHP algorithm (Direct Hashing and Pruning)	Restricts the number of candidate itemsets and reduce the transaction database size
An Efficient Algorithm for Mining Association Rules In Large Databases [25].	Savasere et. al., 1995	Partitioning algorithm	This algorithm decreases database scan to generate frequent itemsets Reduces the I/O overhead and CPU
Dynamic Itemset Counting and Implication Rules For Market Basket Data [26].	Brin et al., 1997	DIC(Dynamic itemset counting) algorithm	Decreases the number of transitions through the database and use fewer candidate itemsets than approaches based on sampling.
Sampling Large Databases for Association Rules [27].	Toivonen. 1996.	Sampling Algorithm	Reduces the database activity for finding association rules, less scan or time.
Online Association Rule Mining [28].	Hidber, 1999	CARMA(Continuous Association Rule Mining Algorithm)algorithm	Out-performs Apriori and DIC on low support thresholds. Use memory more efficiently
SAM: A Split And Merge Algorithm For Fuzzy Frequent [29].	Borgelt and Wang, 2009	SAM(Split and Merge Algorithm)	This algorithm can beimplementedon external storage or relational databases easily
PRICES: An Efficient Algorithm For Mining Association Rules [30].	Wang and Tjortjis, 2004	PRICES Algorithm	Reduces large itemset generation time. It is ten times as quick as Apriori in some cases
Vertical Layout Based			
Study	**Authors / Year**	**Method**	**Advantages**
Scalable Algorithms For Association Mining [31].	Zaki, 2000	Six new algorithms combining these features (ECLAT (Equivalence CLAss Transformation), MaxEclat, Clique, MaxClique, TopDown, and AprClique)	MinimizesI/O costs by making only a small number of database scans, decreases computation costs
Projected Layout Based			
Mining Frequent Patterns Without Candidate Generation [33].	Han et al.,2000	FP-growth method	Saves the costly database scans in the subsequent mining processes and decreases the size of the subsequent conditional pattern bases and conditional FP-trees.
HMine: Fast And Space Preserving Frequent Pattern Mining In Large Databases [34].	Pei et al., 2007	H-Mine algorithm	H-mine has an great performance for different kinds of data and a polynomial space complexity.

Two main methods have been developed based on this idea: (1) GSP, a horizontal format-based sequential pattern mining method, (2) and SPADE, a vertical format-based method.

Generalized Sequential Patterns Algorithm (GSP)

GSP is a horizontal data format based sequential pattern mining algorithm proposed by Srikant and Agrawalin 1996. It contains time constraints, a sliding time window, and user-defined taxonomies. In this algorithm, it uses the downward-closure property of sequential patterns and adopts a multiple pass, candidate generate-and-test approach [38].

TABLE 9: SEQUENTIAL PATTERN MINING

Apriori Based			
Study	**Authors / Year**	**Method**	**Advantages**
Mining Sequential Patterns: Generalizations And Performance Improvements [38].	Srikant, Agrawal,1996	Generalized Sequential Patterns (GSP)	GSP is much faster than the Apriori All algorithm. It guarantees finding all rules that have a user-specified minimum support.
SPADE: An Efficient Algorithm For Mining Frequent Sequences [39].	Zaki, 2001	SPADE algorithm	SPADE outperforms the best previous algorithm. Problems can be solved in main memory easily and efficiently. .
Pattern-Growth-based			
FREESPAN: Frequent Pattern-projected Sequential Pattern Mining [40].	Han,2000	FREESPAN (Frequent pattern-projected Sequential Pattern Mining)	FREESPAN mines the complete set of patterns and runs more efficiently and faster than Apriori-based GSP algorithm.
Mining Sequential Patterns By Pattern-growth: The PrefixSpan Approach [41].	Pei et al.,2004	PrefixSpan algorithm	PrefixSpan has better performance than the apriori based algorithm GSP, FREESPAN, and SPADE.

Sequential Pattern Discovery Using Equivalent Classes) Algorithm (SPADE)

SPADE algorithm is an Apriori-Based Vertical Data Format algorithm represented by Zaki (2001).The algorithm decomposes the original problem into smaller sub-problems which can be easily solved in main memory using efficient lattice search techniques and simple join operations [39].

3.2.2. Pattern-Growth-based Approaches

These approaches provide efficient mining of sequential patterns in large sequence databases without candidate generation. Two main Pattern-Growth algorithms are Frequent pattern-projected Sequential Pattern Mining (FREESPAN) [40]and Prefix-projected Sequential Patterns Mining (PrefixSpan) [41].Table 9 shows Sequential Pattern Mining algorithms.

FREESPANAlgorithm

FREESPAN algorithm is proposed by Han et al. in 2000 for the purpose of reducing efforts of candidate subsequence generation. This algorithm uses frequent items to recursively project sequence databases into a set of smaller projected databases. This work showed that FREESPAN mines the complete set of patterns and runs more efficiently and faster than Apriori-based GSP algorithm [40].

PrefixSpan Algorithm

This is a pattern-growth approach to sequential pattern mining, which was developed by Pei et al. in 2001 [41]. Again, PrefixSpan works in a divide-and-conquer way. This algorithm projects recursively a sequence database into a set of smaller projected databases and reduces the number of projected databases using a pseudo projection technique [41].

GSP, SPADE, and PrefixSpan have been compared by Han et al. in 2004. PrefixSpanhas better performance than GSP, FREESPAN, and SPADE and consumes smaller memory space than GSP and SPADE.

3.3. Structured Pattern Mining

Complicated scientific and commercial applications need to resolve more complicated patterns than frequent itemsets and sequential patterns. For example sophisticated patterns consist of trees, lattices, and graphs. Graphs play a major role in modeling sophisticated structures .They are used in various applications such as, chemical informatics, text retrieval, video indexing, bioinformatics, web analysis, and computer vision. Frequent substructures can be discovered in a collection of graphs. Washio and Motoda in 2003 provided a survey on graph-based data mining [42]. Several methods have been developed for mining interesting subgraph patterns from graph datasets such as mathematical graph theory based approaches like FSG and GSPAN, greedy search based approaches like SUBDUE, inductive logic programming (ILP) based approaches like WARMR etc. A short summary of these algorithms are shown in Table 10.

SUBDUE Algorithm

SUBDUE algorithm is a graph-based relational learning system which is developed by Holder et al. in 1994 improved over the years [43].

SUBDUE produces a smaller number of substructures in graph datasets by finding subgraphs and can efficiently discover best compressing frequent patterns. This algorithm is very efficient for finding recurring subgraphs in a single large graph [44].

Frequent SubGraph Discovery Algorithm (FSG)

FSG, algorithm is proposed by Karypis and Kuramochi in 2004 for finding frequently occurring subgraphs in large graph datasets. This algorithm can be used to discover

recurrent patterns in spatial, scientific and relational datasets. This work showed that FSG is efficient for finding all frequently occurring subgraphs in datasets that contain over 200,000 graph transactions and scales [45].

Graph-based Substructure pattern mining Algorithm (GSPAN)

GSPAN algorithm discovers frequent substructures without candidate generation. This algorithm can be used for mining all kinds of frequent substructures including sequences, trees, and lattices. GSPAN algorithm mines frequent subgraphs more efficiently than others [46]. Also, it outperforms FSG algorithm in mining larger frequent subgraphs in a bigger graph set with lower minimum supports.

Inductive Logic Programming Algorithm (WARMR)

WARMR, a powerful Inductive Logic Programming (ILP) Algorithm, was presented by King et al. in 2001.WARMR is the first ILP data mining algorithm to be used to chemoinformatic data. WARMR extends Apriori to discover frequent queries in data by using rules to generate the candidates from frequent queries and mines Association Rules in Multiple Relations (ARMR's) [47].WARMR has a strong advantage over previous algorithms for discovery of frequent patterns.

TABLE 10: STRUCTURED PATTERN MINING

Structured Pattern Mining			
Study	**Authors / Year**	**Method**	**Advantages**
Substructure discovery in the SUBDUE system [43].	Holder et al., 1994	SUBDUE algorithm.	It is very efficient for finding recurring subgraphs in a single large graph.
An Efficient Algorithm for Discovering Frequent Subgraphs [45].	Karypis and Kuramochi, 2004	FSG (Frequent SubGraph discovery) Algorithm:	It is efficient for finding all frequently occurring subgraphs in datasets containing over 200,000 graph transactions and scales.
Graph-Based Substructure Pattern Mining [46].	Yan and Han,2002	GSPAN Algorithm	It mines frequent subgraps more efficiently with lower minimum supports.
WARMR: A Data Mining Tool For Chemical Data [47].	King et al, 2001	WARMR (Inductive Logic Programming) Algorithm.	It has a strong advantage over previous algorithms for discovery of frequent patterns.

4. RESULTS AND CONLUSION

Since the substantial advancements in computer era, mining of big data to gain useful knowledge has been a hot topic studied from several aspects. One of the important fields of research is ARM which aims to uncover the subtle relations between the entries of huge bulk data so that some meaningful rules of associations can be generated. The obtained association rules can be exploited to identify which instances correlate in certain dimensions. ARM techniques have been used in many different areas ranging from retail industry to healthcare and diagnosis of illnesses.

In this work, a comprehensive literature review on the existing algorithms of ARM is conducted with a special focus on the performance and application areas of the algorithms.

Many algorithms have been proposed to discover association rules since the beginning of research in this area. The developed algorithms are, in general, classified into three main classes: (1) based on frequent itemsets, (2) based on sequential pattern, and (3) based on structures pattern.

This classification mainly groups the algorithms based on the given structure of the dataset. Thus, the structure of the dataset subject to ARM study essentially determines the algorithm to be employed. Within each of these three classes, each algorithm provides superiority in certain aspects in comparison with each other. The above discussed algorithms were developed to improve the accuracy and decrease the complexity, and execution time. However, they do not always succeed to optimize all these objectives simultaneously.

REFERENCES

[1] A. Baykal, “Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Bir Veri Seti Üzerinden Karşılaştırılması,” in XVII. Akademik Bilişim Konferansı, Eskisehir, 2015.

[2]V. Gancheva, “Market Basket Analysis Of Beauty Products”, M.S. Thesis, Erasmus University Rotterdam ,Erasmus School of Economics, 94p, Rotterdam, 2013.

[3] F. Ö. Bükey, “Data Mining Applications in Customer Relationship Management And A Comparative” ,Ph.D. Thesis, Department of Industrial Engineering, Marmara University,172p, Istanbul, 2014.

[4] B. Sherdiwala, O. Khanna, “Association Rule Mining: An Overview”, International MultidisciplinaryR esearch Journal (RHIMRJ), 2015.

[5] G. Serban, I. G. Czibula and A. Campan, “A Programming Interface For Medical diagnosis Prediction”,StudiaUniversitatis, "Babes-Bolyai", Informatica, LI(1), pages 21-30, 2006.

[6] N. Gupta, N Mangal, K. Tiwari and P. Mitra, “Mining Quantitative Association Rules in Protein Sequences”, In Proceedings of Australasian Conference on Knowledge Discovery and Data Mining –AUSDM, pages 273-281, 2006.

[7]D. Malerba, F. Esposito and F. A. Lisi,“Mining spatial association rules in census data”, In Proceedings of Joint Conf. on New Techniques and Technologies for Statistics and Exchange of Technology and Know-how, pages 541-550, 2001.

[8] R. S. Chen,R. C. Wu and J. Y. Chen, “Data Mining Application in Customer Relationship Management Of Credit Card Business”, In Proceedings of 29th Annual International Computer Software and Applications Conference (COMPSAC'05), Volume 2, pages 39-40, 2005.

[9]C. Borgelt, "FrequentItem Set Mining", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery",2(6), pages437-456, 2012.

[10]B. Gupta and D. Garg, "A Taxonomy of Classical Frequent Item set Mining Algorithms", International Journal of Computer and Electrical Engineering,, 3(5), pages 695-699, Jan 2011.

[11] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In Proc. Int'l Conf. Very Large Data Bases (VLDB), pages 487–499,Sept 1994.

[12]M. J. Zaki,M. Ogihara, S. Parthasarathy and W.Li, "Parallel Data Mining For Association Rules On Shared-Memory Multi-processors", In Supercomputing, Proceedings of the 1996 ACM/IEEE Conference on, pages 43-43, 1996.

[13] L. Ji., B. Zhang and J. Li, "A New Improvement on Apriori Algorithm", Computational Intelligence and Security, 2006 International Conference on. Vol. 1. IEEE, Nov. 2006.

[14] L. J.Huang, "FP-growth Apriori algorithm's Application in the Design for Individualized Virtual Shop on the Internet", In Machine Learning and Cybernetics, 2007 International Conference , Vol. 7, Hong Kong, pages 3800-3804, Aug.2007.

[15] H. Wu, Z. Lu, L. Pan, R. Xu and W. Jiang, "An improved apriori-based algorithm for association rules mining", In Fuzzy Systems and Knowledge Discovery, Sixth International Conference on,Vol. 2, pages 51-55, Aug 2009.

[16] X. Shao, "The Application of Improved 3D Apriori Three Dimensional Association Rules Algorithm in Reservoir Data Mining", Proceedings of CIS (1), IEEE Computer Society, pages 64-68, Dec. 2009.

[17] V. Sharma and B. M. M. Sufyan , "A Probabilistic Approach To Apriori Algorithm", International Journal of Granular Computing, Rough Sets and Intelligent Systems,2(3), pages 225-243, 2012.

[18] H. Wang , X. Ji, Y. Xue and X. Liu, "Applying Fast-Apriori Algorithm To Design Data Mining Engine", In Proc. of International Conference on System Science, Engineering Design and Manufacturing Informatization, Vol. 1, pages 63-65, Nov 2010.

[19] Z. Zeng, H. Yang and T. Feng."Using HMT and HASH_TREE To Optimize Apriori Algorithm", International Conference on Business Computing and Global Information. IEEE, pages 412-415, 2011.

[20] L. Xiaohui, "Improvement of Apriori algorithm for associationrules", World Automation Congress (WAC), Mexico,pages1-4,Jun. 2012.

[21] S. Prakash and R. M. S. Parvathi,"An enhanced Scaling Apriori for Association Rule Mining Efficiency", European Journal of Scientific Research, vol. 39, pages.257-264, 2010.

[22]S. Kamrul, K. Mohammad and A. Hasnain, " Reverse Apriori Algorithm For Frequent Pattern Mining", Asian Journal of Information Technology, pages 524-530, 2008.

[23] J. S. Park, M. S. Chen and P. S. Yu, "An Effective Hash-based Algorithm For Mining Association Rules", SIGMOD '95 Proceedings of the 1995 ACM SIGMOD international conference on Management of data, New York, pages 175-186, 1995.

[24]E. T. Wang and A.L. Chen, "A Novel Hash-based Approach for Mining Frequent Itemsets over Data Streams Requiring Less Memory Space", Data Mining and Knowledge Discovery, 19(1), pages 132-172, 2009.

[25] A. Savasere, E. Omiecinski and S. Navathe, "An Efficient Algorithm For Mining Association Rules In Large Databases", Proceedings of the 21th International Conference on Very Large Data Bases, pages 432–444, 1995.

[26] S.Brin, R. Motwani, J. D. Ullman and S. Tsur,"Dynamic Itemset Counting And Implication Rules for Market Basket Data.", In ACM SIGMOD Record, 26(2), pages 255-264, Jun. 1997.

[27] H. Toivonen, "Sampling Large Databases For Association Rules",In 22th International Conference on Very Large Databases (VLDB'96),pages 134–145, Bombay,India., 1996.

[28] C. Hidber, "Online association rule mining", In Proc. Of the 1999 ACM SIGMOD International Conference on Management of Data, 28(2), pages 145–156, 1999.

[29] C. Borgelt and X. Wang, "Sa M: A Split and Merge Algorithm for Fuzzy Frequent Item Set Mining", Proc. 13th Int. Fuzzy Systems Association World Congress and 6th Conf. of the European Society for Fuzzy Logic and Technology (IFSA/EUSFLAT'09, Lisbon, Portugal), pages 968-973, 2009.

[30] C. Wang and C Tjortjis, "PRICES: An Efficient Algorithm For Mining Association Rules", in Lecture Notes Computer Science vol. 3177, pages 352-358, 2004.

[31] M. J. Zaki, "Scalable Algorithms For Association Mining",Knowledge and Data Engineering, IEEE Transactions on, 12(3), pages 372-390, 2000.

[32] S. Neelima, N. Satyanarayana and K. P. A. Murthy, "Survey on Approaches for Mining Frequent Itemsets", IOSR Journal of Computer Engineering (IOSRJCE), 2014.

[33] J. Han, J. PeiandY. Yin, "Mining Frequent Patterns without Candidate Generation", In Proc. ACM SIGMOD Intl. Conference on Management of Data, 29(2), pages 1-12, 2000.

[34] J. Pei, J. Han, H.Lu, S. Nishio, S. Tang and D.Yang, "H-Mine: Fast And Space-preserving Frequent Pattern Mining In Large Databases",IIE transactions,39(6), pages 593-605, 2007.

[35] R. C. Agarwal, C. C. Aggarwal and V. V. V. Prasad, "A tree projection algorithm for generation of frequent item Sets",Journal of parallel and Distributed Computing, 61(3), pages350-371, 2001.

[36] J. Liu, K. Wang, L. Tang and J. Han, "Top Down Fp-growth For Association Rule Mining", Springer Berlin Heidelberg, pages 334-340, 2002.

[37] G. Grahne and J. Zhu, "Efficiently Using Prefix-trees in Mining Frequent Itemsets", In IEEE ICDM'03 Workshop FIMI'03, Melbourne, Florida, USA, 2003.

[38] R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations And Performance İmprovements", In Proc. 5th Int. Conf. Extending Database Technology (EDBT'96), Avignon, France, pages 3-17, 1996.

[39] J. M. Zaki." SPADE: An efficient algorithm for mining frequent sequences", Machine Learning, pages 31-60, 2001.

[40] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal and M. C. Hsu, "FREESPAN: Frequent Pattern-projected Sequential Pattern Mining", In Proc. 2000 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'00), Boston, MA, pages 355-359, 2000.

[41] J. Pei,J. Han, J. Mortazavi- Asl, H. Pinto, Q. Chen, U. Dayal, and M. C. Hsu, "Prefix Span: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", 17th International Conference on Data Engineering (ICDE), 2001.

[42] T. Washio and H. Motoda," State Of The Art Of Graph Based Data Mining", SIGKDD Explor 5, pages 59–68, 2003.

[43]L. B. Holder, D. J. Cook and S. Djoko, "Substructure Discovery In The Subdue System", In: Proceeding of the AAAI'94 workshop knowledge discovery in databases
(KDD'94), Seattle, WA, pages 169–180, 1994.

[44] N. S. Ketkar, L. B. Holder and D. J. Cook, "Subdue: Compression-Based Frequent Pattern Discovery in Graph Data", OSDM '05,pages 71-76, 2005.

[45] M. Kuramochi and G. Karypis, "An Efficient Algorithm for Discovering Frequent Subgraphs", IEEE Trans. Knowl. Data Eng.16(9), pages 1038-1051, 2004.

[46] X. Yan and J. Han, "GSPAN: Graph-Based Substructure Pattern Mining", In ICDM'02: 2nd IEEE Conf. Data Mining, pages 721–724, 2002.

[47] R. King, A. Srinivasan and L. Dehaspe, "WARMR: A Data Mining Tool For Chemical Data", Journal of Computer-Aided Molecular Design 15, pages 173–181, 2001.

[48] E. Duneja and A. K. Sachan, "A Survey on Frequent Itemset Mining with Association Rules", International Journal of Computer Applications, 46(23), pages 18-24, 2012.

[49] D. Palagin, "Mining Quantitative Association Rules in Practice", School of Computer Science and Communication, Royal Institute of Technology M.Sc. Thesis, 73p, Stockholm.
2009.

[50]J. Han, J. Pei and X. Feng, " From Sequential Pattern Mining to Structured Pattern Mining: A Pattern-Growth Approach", Yan University of Illinois at Urbana-Champaign, Urbana, IL 61801, U.S.A. State University of New York at Buffalo 19(3), pages 257-279, 2004.