Modern Applications and Challenges for Rare Itemset Mining

Sadeq Darrab, David Broneske, and Gunter Saake

Abstract—**Data mining is the process of extracting useful unknown knowledge from large datasets. Frequent itemset mining is the fundamental task of data mining that aims at discovering interesting itemsets that frequently appear together in a dataset. However, mining infrequent (rare) itemsets may be more interesting in many real-life applications such as predicting telecommunication equipment failures, genetics, medical diagnosis, or anomaly detection. In this paper, we survey up-to-date methods of rare itemset mining. The main goal of this survey is to provide a comprehensive overview of the state-of-the-art algorithms of rare itemset mining and its applications. The main contributions of this survey can be summarized as follows. In the first part, we define the task of rare itemset mining by explaining key concepts and terminology, motivation examples, and comparisons with underlying concepts. Then, we highlight the state-of-art methods for rare itemsets mining. Furthermore, we present variations of the task of rare itemset mining to discuss limitations of traditional rare itemset mining algorithms. After that, we highlight the fundamental applications of rare itemset mining. In the last, we point out research opportunities and challenges for rare itemset mining for future research.**

Index Terms—**Data mining, rare itemsets, survey.**

I. Introduction

Data mining is a process for discovering hidden knowledge from data. Extracted knowledge should be interpretable by humans in order to support easy decision-making. The fundamental tasks of data mining can be classified into three parts: clustering, classification, and association rule mining [1]. While there is a plethora of work for clustering and classification, association rule mining (ARM) is still an active research topic. ARM indicates the strong relationship between groups of items that frequently appear together in a dataset. For example, in a retail store dataset, the association {beer → chips} indicates that if customers buy beer, they often buy chips as well. By this association, managers can take decisions to increase sales by offering discounts. Since it is emerged in 1990s by Agrawal *et al.* [2], ARM has gained a lot of attention due to its necessity in many domains, such as bioinformatics, image classification, network traffic analysis, medical diagnosis, and market basket analysis [3]. ARM requires two essential steps: 1) frequent itemset mining (FIM) in which all frequent itemsets are generated, and 2) from the previously generated patterns, all ARMs are discovered. Most of the studies [2], [4]-[6] focus on the first step since generating frequent itemsets is computationally expensive whereas generating association rules is straightforward.

FIM aims to discover itemsets that should co-occur together in a given dataset. It focuses on discovering useful information that can be represented as patterns. The resulting patterns can be understandable by humans and hence help experts to take right decisions. For example, discovering patterns of frequently purchased groups of items in basket market leads to a better understanding the customer's behavior. In order to shed some light into this area, there are several recent extensive studies for mining frequent itemsets [2]-[6]. However, the generated frequent itemsets represent widely-known phenomena and are, thus, usually unsurprising and predictable. Hence, analyzing the data for more interesting and unexpected patterns is desirable to uncover hidden knowledge from the data. To this end, detecting abnormal events came into focus, which is known as rare itemset mining [7]-[15].

First attempts to support rare itemsets for traditional approaches fail due to the following two problems. First, several traditional methods were extended to mine rare itemsets by lowering their minimum support in order to still include interesting patterns. However, this resulted in a massive amount of patterns being generated and hence they are computationally expensive to analyze. Second, to overcome this issue, a typical approach for traditional methods is to limit discovered patterns with a high minimum support threshold. This, however, may result in the failure to discover useful itemsets in the dataset since the interesting rare (infrequent) itemsets may be ignored. This problem is called rare item problem, which is still not fully solved [7].

In summary, mining rare itemsets represents the situations where experts have a high interest of abnormal events revealing valuable information from items whose support is low. Recently, there are several studies presented to discover rare itemsets due to its importance in many real-life applications such as medical research, DNA gene and homeland security. For example, mining rare itemsets of passenger's behavior in airport can help us to identify suspicious and unknown threats [16].

In the field of rare itemset mining, there is a survey presented by Koh *et al.* [17], which gives a good overview of rare itemset mining. However, this survey is no longer up-to-date, because the state of the art has changed since then. Furthermore, mining rare itemsets with multiple minimum support is not extensively presented and advanced topics, such as big data challenges in mining rare itemset, has not been discussed. To fill this gap, our goal in this paper is to

Manuscript received November 12, 2019; revised September 26, 2020.

Sadeq Darrab is with the University of Magdeburg, Magdeburg, Germany (e-mail: Sadeq.darrab@ ovgu.de).

David Broneske and Gunter Saake are with the Databases and Software Engineering Group of Gunter Saake, University of Magdeburg, Magdeburg, Germany (e-mail: dbronesk@iti.cs.uni-magdeburg.de, saake@iti.cs.uni-magdeburg.de).

doi: 10.18178/ijmlc.2021.11.3.1037

present a comprehensive up-to-date survey of rare itemset mining. The main contributions in this survey are:

1) In-depth understanding of rare itemset mining is introduced by explaining concepts, motivation examples, and comparisons with underlying concepts,
2) The state-of-art methods for mining rare itemsets are highlighted.
3) Various extensions of existing rare itemset mining are introduced.
4) Applications of rare itemsets are highlighted.
5) Fundamental problems and research opportunities for rare itemset mining are pointed out.

The rest of this paper is organized as follows. In Section II, we describe the background and problem description of both frequent and rare itemset mining, in addition to the relevant definitions employed in the itemset mining. In Sections III and IV, we discuss the most common approaches for mining rare and frequent itemsets including rare ones, respectively. Afterwards, in Section V, we provide various itemset mining problems related to rare itemset mining. In Section VI, we suggest the application domains where rare itemset mining methods can be used. At last, we highlight open challenges and opportunities in rare itemsets mining in Section VII and outline conclusions in Section VIII.

II. Background and Problem Description

In this section, frequent itemset mining task and its related concepts are presented. We discuss the basic concepts of frequent pattern mining for the discovery of frequent itemsets, rare itemset, and interesting associations between itemsets in transactional datasets. We will use the following motivating example to illustrate the concepts of frequent and rare itemsets.

Motivating example: Given a transaction dataset as in Table I, a minimum support threshold = 50%, and minimum confidence threshold = 80%, FIM's task is to find the complete set frequent itemsets. The MIS values of items are given Table II.

TABLE I: Transaction Dataset

TID	Items
1	a, c, d
2	a, c, e
3	a, b, c, e
4	b, e
5	a, b, c, e

TABLE II: MIS Values of Items

Item	a	b	c	d	e
MIS	60%	70%	60%	40%	100%
Support	80%	60%	80%	20%	80%

A. *Problem Definition*

The aim of FIM is to discover useful knowledge from a given dataset. This knowledge can be represented as patterns (itemsets). A pattern is a collection of items that co-occur together. The problem of FIM is formally defined by Agrawal *et al.* [2] as follows.

Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n items and a transaction dataset DB = $\{T_1, T_2, \ldots, T_m\}$ represents a set of m transactions. Each transaction T_p has a unique identifier called transaction identifier (TID) where $T_p \subseteq I$ $(1 \leq p \leq m)$. An itemset $X \subseteq I$ is a set of items and |X| denote to the length of the itemset X as the number of items in X. The itemset with k items is called k-itemsets. For instance, in our motivating example an itemset X= {ace} is called 3-itemset since it contains three items $\{a, c, e\}$. The number of transactions that contain an itemset X is known as the support of X. For example, the support of the itemset {ace: 3} is 3 since it occurs in transactions 2, 3, and 5.

Definition 1. Frequent itemset: An Itemset X is called frequent if its occurrence frequency (Sup) in a given dataset DB no less than the user-given minimum support threshold, minsup, such that Sup (X) > = minsup. For example, in our motivating example, the itemset {ace: 3} is frequent since its support = 60% satisfies the given minimal support of minsup = 50% [2].

Property 1. Downward closure property: Any subset of frequent itemsets must be frequent. For instance, in our motivating example, the itemset {ace: 3} is frequent. Therefore, all its subsets {ac: 3, ae: 3, ce3, a: 4, e: 4. c: 4} are frequent according to Property 1 [2].

Definition 2. Association rule: An association rule $X \rightarrow Y$ means that the occurrence of X implies the occurrence of Y. This rule is called strong if it satisfies two measures as follows [2].

1) The support measure (Sup ($X \rightarrow Y$)) represents the occurrence frequency of transactions in the DB that contain X and Y such that Sup ($X \rightarrow Y$) = Sup ($X \cup Y$) / |DB|, where |DB| refers to the size of a given dataset.
2) The confidence measure (Conf ($X \rightarrow Y$)) refers to the number of transactions in the DB that when containing X also contain Y, such that Conf ($X \rightarrow Y$) = Sup ($Y|X$) = Sup ($X \cup Y$)/Sup(X). Thus, the association rule $X \rightarrow Y$ is interesting if its support and confidence measures satisfy both a minimum support threshold (minsup) and a minimum confidence threshold (minconf). For example, the rule a → c is strong since its support = Sup ($a \cup c$) = 4/5 = 0.80 ≥ minsup= 50%, and its confidence Conf ($a \rightarrow c$) = 80/80 = 100%.

Definition 3. Rare itemset: An itemset X is called rare itemset if its occurrence frequency in a given dataset DB less than the user-given minimum support threshold, minsup, such that Sup (X) < minsup. For example, in our motivating example, the itemset {ab: 2} is rare itemset since its support = 40%, which is less than minsup = 60% [8], [18].

Definition 4. Multiple Item Support (MIS): Each item $i \in I$ has a different minimum item support (MIS). The minimum items support of an itemset X with k-items is defined as MIS(X) = MIN {MIS($x_1, x_2, \ldots, x_k$). For example, MIS of an itemset {ace} = min {MIS(a), MIS(c), MIS(e)} = 60%. Thus, the itemset {ace} is frequent as its support = 80% ≥ 60% [11], [18].

Definition 5. Frequent closed itemset: An itemset X is called closed in a given dataset DB if none of its proper super itemset Y has the same occurrence frequency such as Sup (X) = Sup(Y). The closed itemset X is called frequent closed itemset if Sup (X) ≥ minsup. For example, an itemset {ac: 4} is not a frequent closed itemset since it has a proper super itemset {ace: 4} with the same support as {ac:4} as shown in Fig. 1 [19].

Definition 6. Maximal frequent itemset: An itemset X is a maximal frequent itemset in a given dataset DB if X is frequent, and there is no super itemset Y of X such that $X \subset Y$ and Y is frequent in DB. For example, an itemset {ace :4} is maximal frequent itemset as there is none of its super itemset is frequent as shown in Fig. 1 [20].

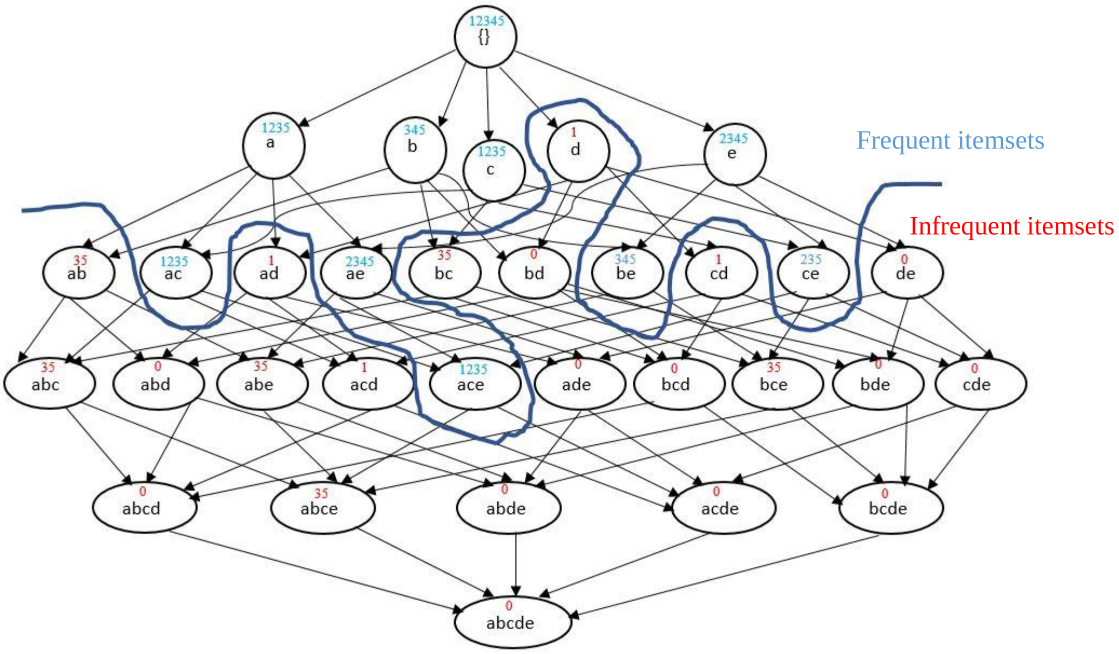

Fig. 1. The lattice represents search space for $I = \{a, b, c, d, e\}$. Each circle in the lattice is an itemset where the number above each itemset is the count of this itemsets for the given dataset in Table I. Itemsets above the borders are frequent itemsets (blue ones) with minsup = 3 whereas itemsets under the border are called infrequent (with red colors).

B. Frequent Itemsets Mining

Frequent itemset mining aims at discovering the whole set of interesting patterns that satisfy a user-defined minimum support threshold. The first attempt to discover the whole set of frequent itemsets is to generate all possible itemsets. Then, those itemsets that meet the user-specified minimum support are considered interesting. This na ïve approach is inefficient since with n items in a dataset, 2^n -1 possible itemsets will be generated. For example, mining frequent itemsets from a dataset that contains 100 items generates $2^{100} - 1$ candidate itemsets. Discovering the interesting patterns from this amount of frequent itemsets is expensive in term of runtime and memory consumption.

To overcome this problem and shrink the search space, Apriori principle in Property 1 has been proposed by Agrawal Rakesh, and Ramakrishnan Srikant. [2] to prune itemsets that contain at least one infrequent subset itemset (item). Subsequently, many approaches have been proposed in order to shrink the search space and to improve the performance. Most of the previously proposed algorithms can be grouped to Apriori-based methods [2], [21], [22] and FP-growth-based methods [23]-[25]. Apriori algorithm and its optimizations methods [2], [3], [8] are based on the Apriori principle technique to shrink the search space.

These methods determine the candidate k+1-itemsets from the frequent k-itemsets. Then, the dataset is scanned iteratively for each candidate itemsets to test whether the generated candidate itemsets satisfy minsup or not. The process is repeated until no candidate itemsets can be generated. These methods show a good performance by utilizing Property 1 to reduce the size of candidates. Nevertheless, Apriori-based methods are expensive as they scan the dataset many times and check a large set of candidates by employing a candidate-set-test-generation approach. To avoid scanning the database iteratively as in Apriori-based algorithms, a vertical representation of dataset called Tid-set or diffsets data structure is presented by Zaki Mohammed J., and Karam Gouda [5], [26], [27]. Using this structure enables us to count the support of a candidate itemset via set intersection instead of scanning the whole dataset iteratively. For candidate-set-test-generation problem, methods that are based on the FP-growth heuristic are introduced [23]-[25]. They use a condensed data structure called FP-tree, to hold all transactions of a given DB in memory. To extract the complete set of frequent itemsets, they recursively mine itemsets using a divide-and-conquer strategy. Although FP-growth-based methods are more efficient as compared to Apriori-based methods, they become inefficient when datasets are sparse. This is because FP-trees become very complex and larger and they cannot fit in the memory if the dataset contains little redundancy and interdependencies.

The above algorithms share the same aim to shrink the search space and increase the performance in term of runtime and memory consumption. They discover the whole set of frequent patterns based on a single minimum support threshold. However, mining with single minsup leads to two problems. The first problem is that mining with low minsup generates a large number of redundant frequent itemsets making them difficult to be analyzed. For mining the non-redundant frequent itemsets, a condensed representation of all frequent itemsets are mined in the form of frequent closed itemsets (FCIs) [28], [29], or maximal frequent itemsets (MFI) [30], [31]. The second problem is that mining with high minsup may result in losing the most interesting patterns called rare itemsets. Mining interesting rare itemsets is a big challenge and it will be discussed in the next suctions.

III. Rare Itemsets Mining

The limitation of FIM algorithms can be described as follows. Setting up a proper minimum support threshold is a critical challenge since with low minimum support, a huge amount of patterns will be generated that are computationally expensive to interpret. In contrary, mining with a high minimum support, the interesting rare (infrequent) itemsets may be missed. Another limitation of traditional FIM algorithms is that they assume that all items have the same nature whereas this is not accurate in real-life applications where some items (frequent itemsets) often occur compared to others (rare itemsets). For example, bread and saucepan do not have the same selling occurrences in a retail market since saucepan is a rarely bought compared to bread but usually creates more profit. Therefore, we may lose useful rare items by traditional methods. This problem is called rare item problem [11]. In many applications, frequent itemsets are less interesting than infrequent (rare) itemsets since frequent itemsets indicate the known and the estimated associations but rare occurring itemsets propose unpredicted associations, which is more valuable to users in many real-life applications such as security in the airports, predicting telecommunication equipment failures, medical diagnosis, genetics and fraudulent credit card transactions [7], [12], [14]. To solve this problem, there are many studies that have been proposed on mining rare itemsets [7]-[15]. In the following subsections we discuss methods that discover only rare itemsets. These methods can be grouped based on search space traversing into breadth-first and depth-first.

A. Breadth-First Search Methods

A breadth-first search (level-wise) approaches explore the search space of itemsets level by level. For example, they first consider itemsets of length 1 called 1-itemsets, then based on the interesting 1-itemsets, the candidate 2-itemsets are generated and only interesting 2-itemsets will be generated and so on. The process is terminated when no candidate itemsets can be generated. To mine rare itemsets using the level-wise approach, there are several studies have been presented [8], [9], [13], [32]-[35].

Adhasivam Kanimozhi and Tamilarasi Angamuthu [9] proposed a method to mine rare itemsets with two automated support thresholds. These minimum support thresholds represent the average support of all unique items and the average of the lowest support and the largest support of the items denoted as AvgSup and MedianSup respectively. Based on these support thresholds, the interesting itemsets could be clustered into three groups as follows.

1) Most_interesting_Group (MiG) that contains the itemsets with support no less than AvgSup.
2) Somewhat_interesting_Group (SiG) that contains itemsets with support between MedianSup and AvgSup.
3) Rare_interesting_Group (RiG) that maintains itemsets with support less than both of AvgSup and MedianSup.

The interesting rare itemsets can be generated by scanning the itemsets in RiG.

Szathmary Laszlo, Amedeo Napoli and Petko Valtchev [8] proposed two methods, MRG-Exp and ARIMA, to discover rare itemsets. These methods classify itemsets into three groups:

1) Minimal generators (MGs) contain itemsets that have support less than their subsets.
2) Minimal rare generators (MRGs) represent itemsets with non-zero support and all its subsets are frequent.
3) Minimal zero generators (MZGs) represent itemsets with zero support but their subsets all have non-zero support.

MRGs represent a boundary that separates frequent itemsets form rare ones. MRG-Exp method uses items in MRGs to create candidates in bottom-up fashion by utilizing generators in MGs. MRGs is utilized by the second method, called ARIMA, to create the whole set of rare itemsets which was created in the first method. The mining process is terminated when MZGs reaches to border since above that boundary there are only no rare itemsets.

Koh Yun Sing, and Nathan Rountree [13] introduced a level-wise algorithm called apriori-inverse to discover rare itemsets. The apriori-inverse is an extended version of the apriori algorithm that utilizes two support thresholds, maximum support (max_sup) and minimum support (min_sup), to discover rare itemsets. The itemsets are considered interesting rare itemsets if their support satisfy min_sup and their support is less than max_sup.

The above mentioned algorithms explore the search space in bottom-up manner. This is expensive because the desirable rare itemsets are usually found at the top. To address this issue, Adda M, Wu L and Feng Y. [33] proposed an apriori-like algorithm called, AfRIM, to recover rare itemsets. This algorithm employs a different method by traversing a lattice top-down starting from the largest itemsets that contain all items in a given dataset. However, it consumes a huge amount of time as it considers the zero-itemsets in the mining process as it starts from the whole set of items in the dataset. To overcome this problem, Troiano *et al.* [32] proposed the rarity algorithm for discovering rare itemset using top-down approach by starting from items in a long transaction. In this algorithm, the borderline is used to separate infrequent itemsets from frequent ones. The itemsets above the borderline are considered rare itemsets.

Tummala, Oswald, and Sivaselvan [34] proposed an algorithm to mine rare itemsets by utilizing a clustering technique. The main contribution is that maximal frequent patterns (patterns with none of its supersets is frequent) and minimal rare itemsets (itemsets that do not satisfy the given minsup but all its subset itemsets are frequent) was separately found by using traditional methods. Then, k of the resulted patterns are used as centroid center for the clustering process. The generated patterns will be added to the clusters based on similarity using Jaccard similarity. Thus, rare itemsets can be founds from specific clusters.

To deal with dramatically growing data, Padillo Francisco, José María Luna and Sebastián Ventura [35] presented a MapReduce-based algorithm called Apriori-Inverse-MR to mine rare itemsets from big data. It involves two steps: 1) it creates rare itemsets whose support are lower than a maximum minimum support but higher than a lower minimum support and 2) it mines perfect association rules that satisfy a minimum confidence threshold by utilizing previously generated itemsets generated in the first step.

B. Depth-First Search Methods

The above methods use a breadth-search approach to discover rare itemsets. However, Apriori-based algorithms have performance drawbacks. They perform multiple scans over the dataset to count the support of candidate itemsets and also they employ a test-generate approach to generate candidate itemsets. To address these drawbacks, a depth-first approach [4] has been introduced to avoid both of multiple scans of datasets and generate-and-test technique.

Tsang Sidney, Yun Sing Koh, and Gillian Dobbie [10] presented a depth-first search algorithm called RP-Tree to mine rare itemsets using a tree structure. It is the first rare association rule mining algorithm that uses a tree structure to mine rare itemsets. It utilizes the FP-tree structure [4] to hold information needed for the mining process. Similar to the FP-growth algorithm, it scans a dataset twice to find rare itemsets. In the first scan, it counts the support of items whereas in the second one it constructs an initial tree, RP-tree. The RP-tree maintains only transactions that contain at least one rare item. Thus, it discards all transactions that contain non-rare items. Thus, mining from RP-tree generates all rare itemsets. An improved and extended version of the RP-Tree, called MCRP-Tree, was proposed by Bhatt and Pratik [7] to mine rare itemsets. The main aim of this method is to find rare itemsets using a tree structure with a maximum constraint model. In this model, different MIS values were assigned to different items and an interesting itemset is discovered only if its support satisfies a maximal MIS value of all items contained in it. Hence, the MCRP-Tree holds all transactions that have at least one rare item. In addition, this method discards all frequent items from the transactions. Thus, it focuses on rare itemsets which give interesting rules and do not spend time in finding uninteresting rules. The MCRP-Tree contains only rare items by discarding all the transactions that do not have rare items.

To handle the dramatic growing of data so called big data, some of efforts have been introduced in [36], [37] by utilizing the power of big data computing platform spark [38]. Liu Sainan, and Haoan Pan [36] proposed a spark-based algorithm to discover rare itemsets from big data. This approach is based on the RP-tree [10] to hold all information needed for the mining process. To overcome apriori-based algorithms drawbacks, the dataset is divided into frequent vertical datasets and rare vertical datasets. These datasets are used to count the support of candidate itemsets. Similar to the RP-tree algorithm, those combinations that do not contain any rare itemsets are discarded. Another method which also based on the spark called SRAM has been proposed to mine abnormal events in wireless network by Liu Ruilin, Kai Yang, Yanjia Sun, Tao Quan, and Jin Yang [37]. This method analyzes network performance counters (NPCs) to identify the root causes of key quality indicator (KQIs) degradation. Thus, discovering the relationship between NPC and KQI outliers is important to identify the root causes of abnormal events in the wireless network by utilizing rare itemsets.

IV. Mining Frequent Patterns Including Rare Itemsets

The limitation of the above traditional FIM algorithms is that they use at most two minimum support thresholds to mine the whole set of rare itemsets. But, items have different nature in real-life applications where some items occur more than others. To overcome this issue, there is a lot of researches in the area of mining frequent itemsets including rare ones [39]-[51]. These methods address the rare item problem by assigning a different minimum support threshold for each item to reflect the nature of each item in a dataset. These methods find out rules that successfully identify the correlations among rare and most occurring items since a user could assign lower minimum supports for the rare (infrequent) items than those for the frequent ones. Thus, these methods discover the interesting itemsets (frequent and rare) whose support is greater than or equal to lowest MIS of its items.

Algorithms to detect frequent itemsets containing rare itemsets can be classified into breadth-first search and depth-first search. In the following subsections, we first discuss breadth-first search and then depth-first search algorithms.

A. Breadth-First Search Methods

There is only a single breadth-first search algorithm for mining frequent itemsets including rare ones under multiple minimum support thresholds called MSapriori which is proposed by Lui Chung-Leunga and Fu-Lai Chung [18]. This method utilizes the original apriori algorithm [52] to discover frequent itemsets involving rare itemsets by assigning a unique minimum item support value for each item. First, MSapriori generates candidate 1-itemsets and then it considers only those items whose support no less than its minimum support value. Secondly, candidate 2-itemsets are generated from 1-frequent itemsets and it discovers 2-frequent itemsets including rare ones who support satisfy the lowest value of its representative items. This approach is terminated when there is no candidate itemsets can be generated. Thus, the generated itemset is interesting if its support exceeds the lowest MIS value of their respective items as shown in Definition 3.

Several algorithms have been proposed to improve the performance of MSapriori [39]-[42]. In [39], Tiantian Xu, and Xiangjun Dong proposed an MSB_apriori method to find frequent itemsets containing rare ones with MIS based on apriori algorithm. Similar to MSapriori, each item in a dataset has its own MIS determined by the user and each itemset is interesting if it can satisfies the lowest MIS of items within it. Unlike MSapriori which uses different steps from Apriori algorithm to mine frequent patterns with MIS, MSB_apriori uses the same steps involved in basic Apriori to find interesting patterns (frequent, rare). MSB_apriori has two main steps as follows: 1) mine all interesting itemsets P by basic apriori with a single minimum support, 2) select the interesting itemsets from *P* that satisfy the definition of frequent itemsets with multiple minimum supports. So, any itemset *X* (*X*∈*P*) must satisfy sup (*X*) ≥ MIS (1) to be frequent since items are in ascending order according their MIS.

In Uday and Krishna [40], an improved multiple support apriori, IMSApriori, algorithm has introduced to discover the interesting itemsets including rare itemsets with MIS. Unlike MSapriori algorithm, it depends on a support difference (SD) to assign the minimum support to items. The SD refers to the acceptable deviation of an item’s frequency so that its itemset can still be considered as interesting. Equation (1) is used to assign $MIS(i)$ for each item.

$$MIS(i) = \begin{cases} S(i) - SD & if(S(i) - SD) > LS \\ LS & \text{otherwise} \end{cases} \tag{1}$$

where, $S(i)$ stands for the support of item i and LS refers to the given lowest minimum support.

The support difference *SD* can *be* calculated as in following equation:

$$SD = \lambda\,(1-\alpha) \tag{2}$$

where λ represents the parameter maximum support of the item supports and α is the parameter ranging between 0 to 1 and SD values changed from $(0, \lambda)$. The authors argue that the SD efficiently decreases the explosion of frequent itemsets including rare items. In [41], [42], two algorithms have been introduced to mine both of frequent and rare itemsets with MIS. They extend the representative algorithm, Apriori, with the following differences. Lee Yeong-Chyi, Tzung-Pei Hong, and Wen-Yang Lin [41] proposed a multiple min-supports mining algorithm to discover all frequent 1-itemsets (including rare ones) for the given dataset by comparing the support of each item with its predefined minimal support. Then, for each k-itemset to be an interesting itemset (frequent and rare), its support is compared with the maximum value of the minimum supports of the items contained in it. Bansal Anubha, Neelima Baghel, and Shruti Tiwari, [42] introduced another method to mine both of frequent and rare itemsets. In this method, all the steps that are used are the same as that used in MSapriori with following exception. The minimum item support thresholds for each item, *i*, can be calculated as in (3).

$$MIS(i) = \begin{cases} \beta\, s(i) & \beta s(i) > LS \\ s(i) & \text{else} \end{cases} \tag{3}$$

where β, $s(i)$ and LS stand for a user-specified value which can be ranged from 0 to 1, support of an item (i), and the least minimum support threshold respectively .

Equation (3) uses to ensure the extraction of frequent itemsets involving rare itemsets. Furthermore, it prunes frequent itemsets involving frequent items in a more efficient manner and without missing the frequent itemsets involving rare items.

The above methods has addresses the rare item problem by discovering both of frequent and rare itemsets. However, they adopt a breadth-first search by utilizing the candidate set generation-and-test method and it is always expensive in terms of execution time and memory consumption, especially for long patterns.

B. Depth-First Search Methods

To avoid limitations of apriori-like algorithms, depth-first search algorithms have been proposed [43]-[51]. A novel multiple item support tree (MIS-Tree) structure has been presented by Hu Ya-Han and Yen-Liang Chen [43] to overcome the problem caused by generate-and-test methods as Apriori-based methods do. Sinthuja, Sheeba Rachel, and Janani [45] introduced the importance of MIS-tree. The authors show that the MIS-tree stores the crucial information about frequent itemsets containing rare ones. Thus, the MIS-tree significantly avoids multiple scans on the dataset by using a generate-test method as Apriori-based algorithms (such as MSapriori) do. In concert with the MIS-tree, a mining algorithm called CFP-growth has been proposed to store the whole information needed to mine frequent and rare itemsets. The CFP-growth algorithm is an extended version of the FP- growth algorithm [4] with the following difference. It scans the dataset once to build the MIS-tree, then useless items are discarded by a reconstruction of the tree. To reconstruct the tree, it employs pruning and merging operations to discard meaningless items from the tree. After the compact MIS-tree is built, the mining process is run to extract the complete set of frequent and rare itemsets. However, CFP-growth has the following limitations, 1) some useless items have been used to build the compact MIS-Tree although they do not generate interesting itemsets (frequent and rare), and 2) these useless items are examined by CFP-growth to build suffix patterns until their respective conditional pattern is empty while no interesting patterns are generated [44].

To tackle these issues, an improved CFP-growth method called CFP-growth++ has been proposed by Kiran *et al*. [44]. CFP-growth++ works similar to CFP-growth with the following differences. The first one is that CFP-growth++ uses the first frequent item for pruning operation, called the lowest minimum support(abbreviation as, LMS) whereas CFP-growth uses the lowest minimum item support MIN. The second difference is that CFP-growth++ introduces a technique to remove a leaf node that cannot generate any frequent or rare itemsets. The last one is that CFP-growth++ utilizes a conditional closure property by which this algorithm discovers all frequent itemsets including rare ones without carrying out mining process till a conditional pattern base becomes empty as CFP-growth does. Thus, execution time and search space can be reduced by CFP-growth++.

There are also some studies [17], [45]-[49] to extract frequent and rare itemsets with multiple minimum support thresholds based on FP-growth method. Taktak Wiem, and Yahya Slimani, [46] presented another multi-support method called MS-FP-growth. Unlike previous methods, MS-FP-growth uses a different minimum support for each level instead of using one minsup value for all items. To discover interesting itemsets, the minsup is varied by increasing and decreasing its value for each level (K-itemsets). Hence, we can discover rare itemsets in the specific level just by using a low minsup. The minsup value can be changed by increasing the value from one level to another, decrease it from one level to another or it can be varied randomly.

Chen Yi-Chun, Grace Lin, Ya-Hui Chan, and Meng-Jung Shih, [17] introduced an algorithm that utilizes the central limit theorem [53] to identify MIS values of items [17]. This method works similar to CFP-growth++ except it uses an automatic tuning MIS method to assign MIS values of items as in (4).

$$\mu(i_j) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{sup}(i_j)$$

$$MIS(i_j) = \mu(i_j) - \sqrt{\tfrac{1}{n}\textstyle\sum_{j=1}^{n}\left(\mathrm{sup}(i_j) - \mu(i_j)\right)^2} \tag{4}$$

where $\mu(i_j)$ refers to the average frequency of the items $\{i_1, \ldots, i_n\}$ that belongs to the same level nodes of each category.

Darrab Sadeq, and Belgin Ergenç [51], [54] proposed MISFP-growth algorithm to mine both frequent and rare itemsets under multiple minimum support. This algorithm is an extended version of FP-growth algorithm. It utilizes MISFP-tree to hold all useful information for mining process without reconstructing the tree as CFP-growth and CFP-growth++ do. It builds the tree with meaningful items that can be considered while mining both frequent and rare itemsets. In [54], MISeclat algorithm has addressed the rare itemset problem by mining frequent patterns including rare ones. It utilizes the TID-list structure to count the support to the generated patterns without scanning the datasets again. To reflect the nature of each items in the given dataset, this algorithm uses different minimum support for different item thresholds for each item. Gan Wensheng, Jerry Chun-Wei Lin, Philippe Fournier-Viger, Han-Chieh Chao, and Justin Zhan [50] proposed a new algorithm named FP-ME to avoid rare itemset problem with MIS. In this method, the ME-tree is utilized to hold all necessary information for the mining process. ME-tree is built with promising items that have support no less than least minimum support, LMS. It avoids building a conditional pattern trees while mining FPs by spanning the ME-tree in depth first without multiple database scans. In addition, they use TID-list and DiffSet data structures to count the support of resulted interesting patterns.

V. Rare Itemset Mining Problems

There are some considerable limitations of the classical methods that are designed to discover useful rare itemsets. Our found limitations can be summarized as follows.

1) Focus on frequency: the first limitation is that they consider the frequency of itemset as the only measure factor to mine interesting rare itemsets. However, there are several factors such as quantities, spatial, utilities, weight of items that should be considered while mining interesting rare itemsets.
2) For dynamic datasets: the second limitation is that conventional algorithms are applicable for static datasets and are incapable of dealing with dynamic scenario.
3) Focus on data format: the third limitation is that the traditional methods for mining rare itemsets are suitable for binary datasets. However, huge amount of data are represented by heterogeneous and diverse types.

To address these limitations, various methods have been introduced to deal with these drawbacks. In this section, we address important limitations.

A. High Utility Rare Itemsets Mining

Conventional rare itemset mining algorithms consider the frequency of the itemsets to be the interesting factor. However, in real life applications, other utility factors such as revenue, cost, quantity, or profit of an item should also be considered according to the user preferences. Hence, finding rare itemsets that have high utility according to end users' interests is desired. For this purpose, Chan Raymond, Qiang Yang, and Yi-Dong Shen [55] introduced utility factor as a measure of how useful an itemset is. In many real life applications, the combinations of rare itemsets with high utilities provide very useful insights to the miner. For example, a sales manager may be interested in infrequent itemsets that generate significant profit. Jyothi, Vyas, and Muyeba [56] proposed a high utility rare itemset (HURI) algorithm to discover high profitable rare itemsets. This algorithm involves two main steps; 1) in the first step, the interesting rare itemsets that do not satisfy the maximum support threshold are discovered; 2) in the second step, high utility rare itemsets that have utility value no less than a minimum utility threshold are generated.

Another method called MHU-Growth (multiple item supports with high utility growth) has been introduced by Ryang Heungmo, Unil Yun, and Keun Ryu, [48] to mine rare itemsets by considering multiple minimum supports and utilities. This algorithm uses both MIS and utilities of items to reflect the nature of each item. MHU-Tree (multiple item supports with high utility tree) which is constructed with a single scan of the dataset and it is utilized to hold information of transactions and high utility of items. To reduce the search space and the number of candidates, four pruning conditions were used to reconstruct the tree. Hence, the meaningful high utility itemsets with multiple minimum supports can be generated from MHU-Tree efficiently. To the best of our knowledge, there is few efforts has been proposed for mining rare itemsets by considering utilities of items.

B. Fuzzy Rare Itemset Mining

Classical data mining algorithms restrict the application area to binary association rules. However, in real life applications, users may be interested in quantitative attributes and hence converting this data into fuzzy data has become an important research direction for data mining applications. Kuok Chan Man, Ada Fu, and Man Wong [57] introduced a method to discover quantitative frequent itemsets. In this algorithm, quantitative values are assigned to each item in transactions. Then, fuzzy membership functions are defined for each item to convert these values to nominal values. Thus, from quantitative data, mining fuzzy specific itemsets brought a useful information especially for the purpose of discovering unusual events. For example, detecting learning problems in educational datasets is a challenging task since many attributes are recorded as quantitative dataset. It is usually known that the most learning behaviors of students are normal and easy to extract in educational datasets whereas unusual activity are not easily extracted. Therefore, studies [58], [59] have been proposed to combine fuzzy set theory with data mining to explore rare itemsets from education datasets. Thus, mining fuzzy rare itemsets is necessary to find out the unusual behavior of students.

C. Mining Rare Itemsets from Dynamic Data

Datasets are frequently updated, and hence a frequent itemset may be infrequent and vice versa. Hoque, Debnath, Easmin, and Rashed [60] proposed an algorithm to mine frequent itemsets including rare ones with MIS from incremental datasets. This algorithm utilizes the MIS-Tree maintenance method that is introduced by Hu *et al*. [43] by violating the restrictions that were used in it. In [43], restrictions were stated that all the data items should be the same in the MIS tree after tuning to void rescanning the dataset again when there is a deleted data in the MIS tree. Hoque, Debnath, Easmin, and Rashed [60] state that there is no need to add any kind of restriction by utilizing another data structure called, a deleted_node_info table which maintains all necessary information about the deleted items. Therefore, the deleted_node_info table helps to recover deleted nodes if needed from the MIS-Tree after support tuning by traversing the path according to the prefix and suffix value of data item that are stored in it. Thus, the tree can be kept in correct status without costly rescanning of the aggregated dataset when an incremental dataset is added to the original dataset.

D. Rare Itemset Mining in Data Streams

Nowadays, numerous amount of data are generated from a wide variety of application areas such as telecommunications call data, sensor data, or online auctions and online bookstores. This amount of data is still flowing on a continuous manner. Traditional rare itemset mining algorithms are no longer applicable in continuous environments where transactions may arrive at a very high speed. To cope with the fast evolving stream of data, several optimizations of algorithms have been designed to process transactions as quickly as possible to count an approximate set of interesting itemsets rather than an exact set of frequent itemsets [61]. In data mining, data stream mining is a challenging problem because of an unbounded sequence of data which arrive in a rapid manner. Furthermore, mining rare itemsets from such data is too expensive. Huang David, Yun Sing Koh, and Gillian Dobbie [62] proposed an algorithm to discover rare itemsets from stream of data. This method constructs streaming rare pattern tree (SRP-tree) to hold all the necessary information required for mining process. The SRP-tree involves a single pass through the dataset using a sliding window approach. Due to nature of data streams, arranging the items altogether is not possible. To handle this problem, a data structure called the connection table is built to efficiently search through the tree and keep a track of items in the window in a canonical order. Thus, when the support of an item becomes less than minimum support threshold, the path containing the item generates all subset of infrequent items.

E. Rare Weighted Itemset Mining

Conventional rare itemset mining algorithms deal similar for a customer who purchase 10 bottles of soda and 5 bags of snacks and another who may purchase 4 bottles of soda and 1 bag of snacks. To reflect the nature of each item and to present its local significance within a transaction in a given dataset, Wang, Yang, and Philip [63] proposed a method to assign a weight for each item in the transaction. This algorithm involves two phases: 1) the frequent itemsets are generated similar to traditional methods, 2) then the weight parameter is considered during the rule generation. Resulted rules are called weighted association rule (WAR). WARs provide a useful insight for managers to do more effective target marketing by identifying customers based on their volume of purchases.

Although mining frequent weighted itemsets is useful, discovering rare weighted itemsets is more interesting in real life applications. Recently, mining infrequent weighted itemsets has gained an attention to mine rarely infrequent weighted itemsets. To discover rare weighted itemsets, Cagliero Luca, and Paolo Garza [64] introduced an algorithm to recover infrequent weighted itemsets from transactional weighted data sets. In this method, two measures are utilized during mining process to discover rare weighted itemsets. The first one is called an infrequent weighted itemset minimum measure, IWI-support-min measure, by which the interestingness of an itemset in a given transaction is weighted by the weight of its least interesting item. The second measure is called IWI-support-max measure which helps to discover interesting itemsets that are weighted by the weight of the most interesting item. These two measures are utilized to select a worthwhile subset of interesting rare weighted itemsets.

VI. Applications for Rare Itemset Mining

In data mining, unusual events can be in the different forms such as outliers, infrequent patterns, or abnormalities. In most applications, unusual behavior detection should be bring useful insights to the data miner. In the following, there are some applications where discovering rare itemsets plays a big role in our daily life.

A. Medical Diagnosis Applications

A significant amount of data is collected from different devices such as electrocardiogram, positron emission tomography, diagnoses, or magnetic resonance imaging. The recognition of unusual patterns in such data often reflects disease conditions and provides useful information for experts. Recently, adverse diseases become a major threat worldwide. Discovering the potential risk at an early stage may result in avoiding the occurrence of threatening diseases such as hepatitis and breast cancer. Extracting useful rare association rules have been used for identifying rare significant correlations between the patient attributes and corresponding diseases [65], [66]. Therefore, extracting the rare complications of diseases has not received enough attention in health care domain.

B. Education Systems Applications

Mining rare itemsets from educational datasets is challenging when applying traditional data mining methods. Discovering rare itemsets can be appropriate for using in educational datasets. Romero Cristóbal, José Raúl Romero, Jose María Luna, and Sebastián Ventura [67] proposed a method to discover rare interesting itemsets from student dataset in a Moodle system. In this study, they show that the resulted rules that are generated by rare itemset mining algorithms can help an instructor to detect abnormal student activities in the Moodle system. Since most attributes are registered as quantitative data in education environment, Weng [58] proposed an Apriori-based mining approach called fuzzy apriori rare itemset mining (FARIM), for mining fuzzy specific rare itemsets from educational datasets to detect unusual learning problems.

C. Anomaly Detection Applications

Anomaly events significantly differ from the normal events. To detect abnormal behavior of a system, a low support threshold is used to discover strong itemsets [68]. Mining rare itemsets has used to detect anomaly in many applications such as credit card fraud detection [69], call intrusion detection, or [70] quality of wireless networks [36]. For instance, Liu *et al.* [37] proposed a method called spark-based rare association rule mining (SRAM) to detect the root cause of wireless network anomaly. The SRAM algorithm discovers the relation network performance counter (NPC) and key quality indicators (KQI) outliers. This relation gives a useful information to the network service providers (NSP) to identify the root causes of anomalies in wireless networks. Another method was proposed by Zengyou, Xiaofei Xu, Joshua Zhexue Huang, and Shengchun Deng [71] to detect anomalies by extracting infrequent itemsets from datasets. In this method, a measure called frequent pattern outlier factor (FPOF) has been introduced to discover the outlier transactions. The FPOF measure is utilized by the FindFPOF algorithm to discover outliers by considering transactions that contain less frequent patterns as outliers.

VII. Research Opportunities and Challenges

Considerable algorithms have been presented to discover interesting rare itemsets. Since the problem of mining unusual events play a big role in real-life applications, there are opportunities and challenges that still remain in rare itemsets mining. In this section, we point out some challenges and opportunities that have the potential to become future research in rare itemset mining.

A. Designing Efficient Approaches

As we see in in the survey, rare itemsets mining is computationally expensive in terms of memory consumption and execution time. Although the proposed algorithms show a good performance to discover abnormal events from the dataset, there is still a gap to be filled by proposing novel data structures, introducing new constraints, designing new pruning approaches to shrink the search space, or setting up a proper measures to find out interesting rare itemsets.

In other situations, currently methods for mining rare itemsets with multiple minimum supports extract both of frequent and rare itemsets. Mining rare itemset with multiple minimum support thresholds is a major challenge.

Furthermore, we have entered the era of big data, and hence there is a rapid growth of the volume of data collected in the various fields. The volume of data is grown tremendously and it is beyond the processing ability of commonly used methods. Although a lot of efforts have been made to cope with the high amount of data for mining frequent itemsets, significant challenges still remain to detect rare itemset mining from big data. Therefore, to handle large-scale data, there are open challenging research opportunities to design a parallelized rare itemset mining algorithms by utilizing Big Data frameworks such as MapReduce [72] and Spark [73].

B. Privacy Preservation

The main goal of data mining is to extract valuable-unknown information from datasets. In real-life applications, rare itemset mining brings more valuable insights to decision making in many domains. However, extracted valuable knowledge may reveal sensitive information which seriously threatens privacy if shared without any precautions. To hide this sensitive information, Öztürk and Ergenç [74] proposed a pseudo graph based sanitization (PGBS) algorithm that focuses on concealing sensitive itemsets on a given transactional dataset. The PGBS algorithm hides the user sensitive itemsets by decreasing their supports below user specified multiple sensitive thresholds. This method may not be applicable in rare itemset mining to hide sensitive information since we are interested in infrequent itemsets that have support less that a given minimum support threshold. To best of our knowledge, no methods have been presented to address privacy preservation for rare itemsets mining. Therefore, it is necessary to address the privacy preserving problem for rare itemset mining.

C. Diversity of Data

There are various types of data and it is recorded in different formats as spatial-time [75], sequence [76], or in continuous manner [62]. There is a need to develop rare itemset mining methods that deal with these different data formats by considering quantities, utilities, weight, continuous of data, and dynamic data. Thus, developing methods to mine interesting rare itemsets to deal with different types of datasets is still crucial and challenging.

D. Specific Applications

Rare itemset mining plays a big role in many domains since it reveals unusual activities so-called called rare itemsets. Detecting unusual events from different applications reflects the real life problems. Therefore, there are research opportunities to apply rare itemset mining in our daily problems such as disease diagnosis, crime analysis programs, DNA genes, security, or detecting abnormality in social networks analysis.

VIII. Conclusion

In this survey, there are several challenges and opportunities for further research. To efficiently explore rare itemset mining, designing an efficient scalable methods is a challenging task. This can be done by considering another efficient data structure while mining rare itemsets, mining most interesting rare itemsets, and mining rare itemsets with multiple minimum support thresholds. Furthermore, dealing with the rapid growth of data to recover rare itemsets is a major challenge. Since rare itemset mining brings more valuable insights to decision making in many domains, privacy concerns become a key challenge while mining interesting rare itemsets. Thus, how to develop a method to extracted rare itemsets without revealing sensitive information in an interesting topic. Furthermore, with the emergence of new technologies such as sensor networks, data comes in a continuous manner. Hence, processing data in an online fashion has become a necessity. In the survey, we also pointed out the necessity for applying rare itemset mining in our daily problems such as disease diagnosis, crime analysis programs, and DNA genes.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Sadeq Darrab conducted the research and wrote the paper. David Broneske and Gunter Saake supervised the whole process giving detailed feedback on the research methodology and paper style. All authors had approved the final version.

References

[1] J. W. Han, P. Jian, and M. Kamber, *Data Mining: Concepts and Technique*s, Elsevier, 2011.

[2] A. Rakesh and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. very Large Data Bases*, 1994, vol. 1215, pp. 487-499.

[3] F.-V. Philippe, J. C.-W. Lin, B. Vo, T. T. Chi, J. Zhang, and H. B. Le, "A survey of itemset mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 4, 2017.

[4] J. W. Han, J. Pei, and Y. W. Yin, "Mining frequent patterns without candidate generation," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 1-12, 2000.

[5] Z. M. Javeed, "Scalable algorithms for association mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, no. 3, pp. 372-390, 2000.

[6] J. Pei, J. Han, H. Lu, S. Nishio, S. Tang *et al*., Hyper-structure mining of frequent patterns in large databases," in *Proc. 2001 IEEE Intern. Conf. Data Mining*, San Jose, USA, 29 November - 2 December, 2001, pp. 441-448.

[7] U. Y. Bhatt and P. A. Patel, "An effective approach to mine rare items using Maximum Constraint," in *Proc. 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO)*, 2015, p. 1-6.

[8] S. Laszlo, A. Napoli, and P. Valtchev, "Tow ards rare itemset mining," in *Proc. ICTAI 2007. 19th IEEE International Conference on Tools with Artificial Intelligence*, 2007, vol. 1, pp. 305-312.

[9] K. S. C. Sadhasivam and T. Angamuthu, "Mining rare itemset with automated support thresholds," *Journal of Computer Science*, vol. 7, no. 3, p. 394, 2011.

[10] T. Sidney, Y. S. Koh, and G. Dobbie, "RP-Tree: Rare pattern tree mining," in *Proc. International Conference on Data Warehousing and Knowledge Discovery*, Springer, Berlin, Heidelberg, 2011, pp. 277-288.

[11] B. Liu, H. Wynne, and Y. M. Ma, “Mining association rules with multiple minimum supports," in *Proc. the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 337-341.

[12] G. M. Weiss, “Mining with rarity: A unifying framework," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 7–19, 2004.

[13] K. Y. Sing and N. Rountree, "Finding sporadic rules using apriori-inverse," in *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, Berlin, Heidelberg, 2005, pp. 97-106.

[14] L. Szathmary, P. Valtchev, A. Napoli, and R. Godin, "Efficient vertical mining of minimal rare itemsets," in *Proc. 9th Intern. Conf. Concept Lattices and Their Applications*, Fuengirola, Spain, 11-14 October, 2012, pp. 269-280.

[15] K. Mohak, C. Oswald, and B. Sivaselvan. "A novel rare itemset mining algorithm based on recursive elimination," *Software Engineering*, Springer, Singapore, pp. 221-233, 2019.

[16] T. Luigi and G. Scibelli, "A time-efficient breadth-first level-wise lattice-traversal algorithm to discover rare itemsets," *Data Mining and Knowledge Discovery*, vol. 28, no. 3, pp. 773-807, 2014.

[17] Y. S. Koh and S. R. Ravana, “Unsupervised rare pattern mining: A survey," *ACM Transactions on Knowledge Discovery from Data*, vol. 10, no. 4, p. 45, 2016.

[18] C.-L. Lui and F.-L. Chung, "Discovery of generalized association rules with multiple minimum supports," in *Proc. European Conference on Principles of Data Mining and Knowledge Discovery*, Springer, Berlin, Heidelberg, 2000, pp. 510-515.

[19] P. Nicolas, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules," in *Proc. International Conference on Database Theory*, Springer, Berlin, Heidelberg, 1999, pp. 398-416.

[20] B. Douglas, M. Calimlim, and J. Gehrke, "Mafia: A maximal frequent itemset algorithm for transactional databases," in *Proc. ICDE*, 2001, vol. 1, pp. 443-452.

[21] Y. Jiao, "Research of an improved apriori algorithm in data mining association rules," *International Journal of Computer & Communication Engineering*, vol. 2, no. 1, pp. 25-27, 2013.

[22] S. Jaishree, H. Ram, and J. S. Sodhi, "Improving efficiency of apriori algorithm using transaction reduction," *International Journal of Scientific and Research Publications*, vol. 3, no. 1, pp. 1-4, 2013.

[23] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 1-12, 2000.

[24] W. Ke, L. Tang, J. W. Han, and J. Q. Liu, "Top down fp-growth for association rule mining," in *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, Berlin, Heidelberg, 2002, pp. 334-340.

[25] Z. Wei, H. Z. Liao, and N. Zhao, "Research on the FP growth algorithm about association rule mining," in *Proc. International Seminar on Business and Information Management*, 2008, vol. 1, pp. 315-318.

[26] Z. M. Javeed, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules," in *Proc. 3rd Intl. Conf. on Knowledge Discovery and Data Mining*, 1997, vol. 97, pp. 283-286.

[27] J. Z. Mohammed and K. Gouda, "Fast vertical mining using diffsets," in *Proc. the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2003, pp. 326-335.

[28] H. Tahrima, M. R. Karim, M. Samiullah, and C. F. Ahmed, "An efficient dynamic superset bit-vector approach for mining frequent closed itemsets and their lattice structure," *Expert Systems with Applications*, vol. 67, pp. 252-271, 2017.

[29] D. Youcef, A. Belhadi, F.-V. Philippe, and J. C.-W. Lin, "Fast and effective cluster-based information retrieval using frequent closed itemsets," *Information Sciences*, vol. 453, pp. 154-167, 2018.

[30] G. Karam and M. J. Zaki, "Efficiently mining maximal frequent itemsets," in *Proc.* 2001 *IEEE International Conference on Data Mining*, IEEE, 2001, pp. 163-170.

[31] V. Bay, S. Pham, T. Le, and Z.-H. Deng, "A novel approach for mining maximal frequent patterns," *Expert Systems with Applications*, vol. 73, pp.178-186, 2017.

[32] T. Luigi, G. Scibelli, and C. Birtolo, "A fast algorithm for mining rare itemsets," in *Proc. 2009 Ninth International Conference on Intelligent Systems Design and Applications*, IEEE, 2009, pp. 1149-1155.

[33] M. Adda, L. Wu, and Y. Feng, 2007, "Rare itemset mining," in *Proc. 6th International Conference on Machine Learning and Applications*, Washington, DC, pp 73–80

[34] T. Kuladeep, C. Oswald, and B. Sivaselvan, "A frequent and rare itemset mining approach to transaction clustering," in *Proc. International Conference on Data Science Analytics and Applications,* Springer, Singapore, 2017, pp. 8-18.

[35] P. Francisco, J. M. Luna, and S. Ventura, "Mining perfectly rare itemsets on big data: an approach based on apriori-inverse and mapreduce," in *Proc. International Conference on Intelligent Systems Design and Applications*, Springer, Cham, 2016, pp. 508-518.

[36] L. Sainan and H. Pan, "Rare itemsets mining algorithm based on RP-Tree and spark framework," in *Proc. AIP Conference Proceedings*, 2018, vol. 1967, no. 1, p. 040070.

[37] R. L. Liu, K. Yang, Y. J. Sun, T. Quan, and J. Yang, "Spark-based rare association rule mining for big datasets," in *Proc. 2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 2734-2739.

[38] Z. Matei, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," *Hot Cloud*, vol. 10, no. 10, 2010.

[39] T. T. Xu and X. J. Dong, "Mining frequent patterns with multiple minimum supports using basic Apriori," in *Proc. 2013 Ninth International Conference on Natural Computation (ICNC)*, 2013, pp. 957-961.

[40] K. R. Uday and P. K. Re, "An improved multiple minimum support based approach to mine rare association rules," in *Proc. IEEE Symposium on Computational Intelligence and Data Mining*, 2009, pp. 340-347.

[41] Y.-C. Lee, T.-P. Hong, and W.-Y. Lin, "Mining association rules with multiple minimum supports using maximum constraints," *International Journal of Approximate Reasoning*, vol. 40, no. 1, pp. 44-54, 2005.

[42] B. Anubha, N. Baghel, and S. Tiwari, "An novel approach to mine rare association rules based on multiple minimum support approach," *International Journal of Advanced Electrical and Electronics Engineering*, vol. 10, pp. 75-80, 2013.

[43] Y.-H. Hu and Y.-L. Chen, "Mining association rules with multiple minimum supports: A new mining algorithm and a support tuning mechanism," *Decision Support Systems*, vol. 42, no. 1, pp. 1-24, 2006.

[44] R. U. Kiran and P. K. Reddy, "Novel techniques to reduce search space in multiple minimum supports-based frequent pattern mining algorithms," in *Proc. 14th International Conference on Extending Database Technology*, 2011, pp. 11-20.

[45] M. Sinthuja, S. Sheeba Rachel, and G. Janani, "MIS-tree algorithm for mining association rules with multiple minimum support," *Bonfring International Journal of Data Mining*, pp. 1-5, 2011.

[46] T. Wiem and Y. Slimani, "MS-FP-growth: A multi-support vrsion of FP-growth algorithm," *International Journal of Hybrid Information Technology*, vol. 7, no. 3, pp. 155-166, 2014.

[47] Y.-C. Chen, G. Lin, Y.-H. Chan, and M.-J. Shih, "Mining frequent patterns with multiple item support thresholds in tourism information databases," *Technologies and Applications of Artificial Intelligence*, Springer International Publishing, 2014, pp. 89-98.

[48] R. Heungmo, U. Yun, and K. H. Ryu, "Discovering high utility Itemsets with multiple minimum supports," *Intelligent Data Analysis*, vol. 18, no. 6, pp. 1027-1047, 2014.

[49] F. A. Hoque, M. Debnath, N. Easmin, and K. Rashed, "Frequent pattern mining for multiple minimum supports with support tuning and

tree maintenance on incremental database," *Research Journal of Information Technology*, vol. 3, no. 2, pp. 79-90, 2011.

[50] W. S. Gan and J. C.-W. Lin, F.-V. Philippe, H.-C. Chao, and J. Zhan, "Mining of frequent patterns with multiple minimum supports," *Engineering Applications of Artificial Intelligence*, vol. 60, pp. 83-96, 2017.

[51] D. Sadeq and B. Ergenç, "Frequent pattern mining under multiple support thresholds," *Wseas Transactions on Computer Research*, pp. 1-10, 2016.

[52] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Record*, vol. 22, pp. 207–216, 1993.

[53] J. A. Rice, *Mathematical Statistics and Data Analysis*, 1995.

[54] D. Sadeq and B. Ergenç, "Vertical pattern mining algorithm for multiple support thresholds," *Procedia Computer Science*, vol. 112, pp. 417-426, 2017.

[55] C. Raymond, Q. Yang, and Y.-D. Shen, "Mining high utility itemsets," in *Proc. Third IEEE International Conference on Data Mining*, 2003, pp. 19-26.

[56] P. Jyothi, O. P. Vyas, and M. Muyeba, "Huri–a novel algorithm for mining high utility rare itemsets," *Advances in Computing and Information Technology*, Springer, Berlin, Heidelberg, pp. 531-540, 2013.

[57] K. C. Man, A. Fu, and M. H. Wong, "Mining fuzzy association rules in databases," *ACM SIGMOD Record*, vol. 27, no. 1, pp. 41-46, 1998.

[58] C.-H. Weng, "Mining fuzzy specific rare itemsets for education data," *Knowledge-Based Systems*, vol. 24, no. 5, pp. 697-708, 2011.

[59] C.-T. Chen and K.-Y. Chang, "A study on the rare itemsets of students' learning effectiveness by using fuzzy data mining," in *Proc. 2016 International Conference on Advanced Materials for Science and Engineering*, 2016, pp. 703-706.

[60] F. A. Hoque, M. Debnath, N. Easmin, and K. Rashed. "Frequent pattern mining for multiple minimum supports with support tuning and tree maintenance on incremental database," *Research Journal of Information Technology*, vol. 3, no. 2, pp. 79-90, 2011.

[61] C. Toon, N. Dexters, and B. Goethals, "Mining frequent itemsets in a stream," in *Proc. Seventh IEEE International Conference on Data Mining*, IEEE, 2007, pp. 83-92.

[62] H. David, Y. S. Koh, and G. Dobbie, "Rare pattern mining on data streams," in *Proc. International Conference on Data Warehousing and Knowledge Discovery*, Springer, Berlin, Heidelberg, 2012, pp. 303-314.

[63] W. Wei, J. Yang, and S. Y. Philip, *Efficient Mining of Weighted Association Rules (WAR)*, IBM Thomas J. Watson Research Division, 2000.

[64] C. Luca, and P. Garza, "Infrequent weighted itemset mining using frequent pattern growth," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 4, pp. 903-915, 2014.

[65] B. Anindita and B. Nath. "Identifying risk factors for adverse diseases using dynamic rare association rule mining," *Expert Systems with Applications*, vol. 113, pp. 233-263, 2018.

[66] P. Saeed, D. Delen, T. Liu, and W. Paiva, "Development of a new metric to identify rare patte rns in association analysis: The case of analyzing diabetes complications," *Expert Systems with Applications*, vol. 94, pp. 112-125, 2018.

[67] R. Cristóbal, J. R. Romero, J. M. Luna, and S. Ventura, "Mining rare association rules from e-learning data," *Educational Data Mining*, 2010.

[68] P. N. Tan, M. Steinbach, andV. Kuma, *Introduction to Data Mining*, Addison-Wesley r, 2005.

[69] R. Brause, T. Langsdorf, and M. Hepp. "Neural data mining for credit card fraud detection," in *Proc. 11th International Conference on Tools with Artificial Intelligence*, IEEE, 1999, pp. 103-106.

[70] L. Wenke, S.J. Stolfo, and K. W. Mok, "Adaptive intrusion detection: A data mining approach," *Artificial Intelligence Review*, vol. 14, no. 6, pp. 533-567, 2000.

[71] Z. Y. He, X. F. Xu, J. Z. X. Huang, and S. C. Deng, "A frequent pattern discovery method for outlier detection," in *Proc. International Conference on Web-Age Information Management*, Springer, Berlin, Heidelberg, 2004, pp. 726-732.

[72] D. Jeffrey and S. Ghemawat, "MapReduce: A flexible data processing tool," *Communications of the ACM*, vol. 53, no., pp. 72-771, 2010.

[73] Z. Matei, M. Chowdhury, T. Das *et al.*, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in *Proc. the 9th USENIX on Networked Systems Design and Implementation*, USENIX Association, 2012, p. 2.

[74] A. C. Öztürk and B. Ergenç, "Itemset Hiding under Multiple Sensitive Support Thresholds," in *Proc. 9th International Joint Conference Knowledge Engineering and Knowledge Management, Funchal*, Madeira, Portugal, 1-3 November 2017, SCI Press, pp. 222-231.

[75] C. Riccardo, F. Porto, E. Pacitti, F. Masseglia, and E. S. Ogasawara, "Spatial sequential pattern mining for seismic data," *The Brazilian Symposium on Databases*, pp. 241–246, 2016.

[76] F.-V. Philippe, A. Gomariz, M. Campos, and R. Thomas, "Fast vertical mining of sequential patterns using co-occurrence information," in *Proc. The Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2014, pp. 40–52.

Sadeq Darrab was born in Hajjah, Yemen. He studied computer science at Sana'a University. He got the bachelor of science in 2004. He received his master of science in 2016 from Izmir Institute of Technology in 2016, Turkey.

Currently, he is a PhD student at the University of Magdeburg, Germany. His research focuses on data mining. His major interest is the development of methods for mining frequent itemset mining and rare itemset mining.

David Broneske was born in 1989 in Burg, Germany. He studied computer science at the University of Magdeburg. He got the bachelor of science in 2012, the master of science in 2013, and he earned his PhD (Dr.-Ing.) from the University of Magdeburg in 2019.

Currently, he is a Post-Doc at the Databases and Software Engineering Group of Gunter Saake at the University of Magdeburg. His research interests include hardware-sensitive database operations and data structures. Dr.-Ing. Broneske is a member of the German Gesellschaft für Informatik (GI e.V.) and earned the research award 2017 at Computer Science Faculty of the University of Magdeburg.

Gunter Saake was born in 1960 in Göttingen, Germany. He received his diploma in 1985 and PhD in 1988 from Technical University of Braunschweig. In January 1993, he received the Habilitation degree (venia legendi) for computer science from the Technical University of Braunschweig.

From 1988 to 1989, he was a visiting scientist at the IBM Heidelberg Scientific Center where he joined the Advanced Information Management project and worked on language features and algorithms for sorting and duplicate elimination in nested relational database structures. Since May 1994, Gunter Saake is a full professor for the area databases and information systems at the Otto-von-Guericke-University Magdeburg. His research interests include database integration, tailor-made data management, database management on new hardware, and feature-oriented software product lines.

Prof. Dr. rer. nat. habil. Saake is member of ACM, IEEE Computer Society, GI and of the organization committee of GI AK `Foundations of Information Systems'. Besides being author and co-author of scientific publications, he is author of a book "Object-oriented Modelling of Information Systems", co-author of four lecture books on Database Concepts, a book on Java and Databases, a lecture book on Object Databases and a book on Building efficient applications using Oracle8 (all in German).