DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4408

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

FIDOOP: PARALLEL MINING OF FREQUENT ITEM SETS USING MAPREDUCE

Dr. K.Kavitha and G.Sudha
Assistant Professor, M.Phil Research Scholar
Mother Teresa Women's University, Kodaikanal, India

Abstract—Due to the exponential increase of real-time data monitoring systems, the extraction of frequent item (Frequent item set mining) set from large uncertain database is the challenging task. The existing parallel mining algorithm for frequent item sets includes the limitations in terms of more memory usage and excessive run time even for less amount of data. To overcome this problem, the FiDoop based item set mining algorithm is proposed by using map reduce framework. It is used to improve the performance of load balancing operation in an uncertain database for computing frequent patterns. This system includes data uploading, preprocessing, threshold, find support and confidence, merge and result. Initially, the data is selected from the dataset and uploaded in the server. Afterwards, the preprocessing stage removes columns which contains unwanted entries. The information is analyzed and partitioned to compute threshold value. These data are classified depends on threshold values and the clustering algorithm is used to find high support and confidence values among clusters to discover frequent item. Finally, those frequent item sets are merged to acquire a frequent pattern. The proposed system is mainly developed for improving the accuracy and it is evaluated based on the performance measures of accuracy, memory usage and execution time.

Index Terms—FiDoop, Map Reduce Clustering, Frequency set mining, support and confidence

I. INTRODUCTION

heprocess of extracting information from the huge data set is termed as data mining. Market analysis, fraud detection, customer retention, production control and science exploration are the applications used to extract the information or knowledge. The FiDoop is used for the frequent itemsets mining algorithm in data mining. The FiDoop is implemented with the mechanism which enables in automatic parallelization, load balancing and data distribution for parallel mining of frequent item sets on the large cluster. FiDoop used the map-reduce programming model because to improve the performance of the frequent itemset mining Hadoop clusters in data mining, the FIM is considered as a critical part of data analysis and it is to extract the information from data sets based on frequently occurring events. [1]Frequent Item Sets Mining (FIM) includes the main problem in association rule mining and sequence mining. FIM algorithms are classified into two, such as,

- Apriori schema
- FP-growth schema

Apriori-based algorithms find the frequent itemsets based on the bottom-up method used for generating the candidate itemsets. Apriori is a classic algorithm for using generate- and – test process which generates a large number of candidate itemsets. It repeatedly scans all database and reduces the time for scanning database. There are multiple techniques are used to generate the frequent item sets. The approaches of generating the frequent sets are divided into three techniques, such as,

- Apriori algorithm
- Eclat algorithm
- FP growth algorithm

*Apriori algorithm:*This algorithm is most traditional and essential for mining the frequent item sets. It is used to find the all frequent item sets in given data set. The states "All non-empty item sets of a frequent itemset must be frequent". Which depends on the apriori algorithm of their property. It follows the two stages, such as,

- Generate phase
- Prune phase

The disadvantage of the algorithm is to generate a large number of candidate sets and it required to repeatedly

analyzing the database and validate the huge set of candidates by pattern matching. It is costly for each transaction in the data base to determine the support of the candidate item set.

*Eclat algorithm:*this algorithm depends on first search based algorithm and it used the vertical database layout. Each item is stored in the cover so it is called as the tid list and uses the intersection based approach to compute the support of an item set. It needs the minimum space than apriori if the item sets are the small number. It is suitable [2] for less number of datasets and minimum time for frequent pattern generation than apriori. The disadvantage of the algorithm is large tid-list at that time it takes more space to store candidate set. It needs more time for intersection when Tid list is large.

FP-Growth Algorithm: FP-Growth is an important frequency pattern mining method that generates the frequent item set without candidate generation. It utilizes the tree [1]based structure so it creates the conditional frequent pattern tree and conditional pattern base which satisfy the minimum support. Fp-growth includes the group of concurrent items and it makes two passes,

Pass 1:

- Scan data and count support for each item
- Discard infrequent items
- Sort frequent items in descending order based on their support

Pass 2:

- Reads one transaction at a time and maps to the tree
- Fixed order is used so that path is shared

- Points are maintain between nodes containing same items
- Frequent items are extracted from the list
- It contain some disadvantages
- Fp tree may not fit in main memory
- Execution time is large due to complex compact data structure

The remaining sections of this paper is organized as follows: Section II reviews some of the existing works related to frequent pattern mining and map reduce. Section III provides the detailed description of the overall work of frequent item sets using map reduce. Section IV presents the performance results of the proposed system. Finally, this paper is concluded in Section V.

II. RELATED WORK

This section presents some of the existing works related to FiDoop base map reduce method. Yuan [3]discussed the mining association rules for an apriori algorithm. It included the two bottlenecks, such as,

- Scanning the data base frequently
- Generating a large number of candidate set.

An inherent defects of apriori algorithm provided the some improvement, such as,

- Using the new database mapping way to avoid scanning the data base repeatedly
- Pruning frequent item sets and candidate item sets in order to improve joining efficiency.
- Using overlap strategy to count support to achieve high efficiency

Apriori algorithm was used to reduce the time consumed in transaction scanning for Calders, et al. [4]generation of candidate itemsets and reduced the number of transaction to scan. An improved algorithm was more efficient compared than the original apriori algorithm. It was used to reduce the time consumption. Tsai, et al. [5]surveyed the data mining for the internet of thing. Many data are created or captured by the IoT and it included with highly useful and valuable information. They were clustering: to classify the pattern of all which are unlabeled.

Classification:it is used to classify the pattern which is labeled and unlabeled.

Association rules:it is used to determine the event from input pattern which occurs the some particular order.

Sequential pattern:the goal is used to find the event from input pattern that occurred in some particular order.

The most of the mining technologies were designed for the finding some useful information which is hidden the data from the original objectives were different. The data mining technologies are the IoT and they included the clustering, classification and frequent pattern mining technologies from the perspective of infrastructures and perspective of services. The classification matrix and the three rules are easily determined the mining technology.Yabing [6]proposedthe apriori algorithm in data mining association rules. It is a classical algorithm of association rules that compute all frequent item sets. The association rule mining was used to discover the association and relations among the item sets of large data. This algorithm was used to solve the efficient problem of association rules mining algorithm.

Advantages of the algorithm

- An optimized algorithm was similar with the apriori algorithm.
- It reduce the time for repeated process
- It save the time cost and increase the efficiency of data mining

Nori, et al. [7]introduced an effective algorithm for closed frequent item sets mining that operates the sliding window model. It was used the data structure for storing transactions of the window and corresponding closed item sets. The datasets were named as T_CET for storing and updating the closed frequent item sets of the window. The advantage was smaller in time, and it required low memory. The main drawback was it provide the less accuracy.Lee, et al. [8]proposed an algorithm for mining weighted maximal frequent patterns (WMFPs) over the data streams based on the sliding window model. The data streams were based on the sliding window techniques which was related to the pruning strategies for eliminating the meaningless pattern. WMFP-SW was effectively used to extract the WMFP-tree and WMFP-SW-array. WMFP-SW guaranteed the more outstanding performance in terms of runtime, memory usage, and scalability.Pyun, et al. [9]proposed the new structure of efficient frequent pattern mining based on linear prefix tree. This structure included the drawbacks in run time and memory usage. The LP-growth constructed the LP-tree with the help of the algorithm. The LP-growth was used to apply in the mining process. The main intention of the algorithm is to reduce the memory usage but the time was increased. LP-tree contributed for improving the performance of frequent pattern mining so it used only the less memory and it was accessed in many cases, such as applications. The algorithm was not only used the in general frequency pattern mining it also used in the variety of pattern mining fields such as closed or maximal pattern mining.

Lee, et al. [8]proposed the genetic algorithm for numerical association rule mining based on the notion of rough pattern. A set of values represented the upper and lower intervals of the rough values. Association rule mining was considered as the problem in the multi objective. It included the three efficient rule, such as,

- Confidence
- Comprehensibility
- Interestingness

Multi-objective genetic algorithm association rule mining was used in data mining with databases which was included in numeric attributes. It was used to discover the numerical association rule. D'Oca and Hong [10]the framework included the two data set mining techniques, such as clustering and association rules. The main objective of the method was motivational, opening duration and interactivity and degree of opening position. The four behavioral patterns were used in association rule for the window opening in office user profiles, such as physical environmental driven and contextual driven. When interacting with windows to restore the indoor environmental quality and the environmental parameters were less impact and the user driven by the contextual factor. It provides the better accuracy compared than other method. Fournier-Viger, et al. [11]presented the algorithm for high utility itemset mining named as Fast High Utility Miner (FHM). The algorithm was used to reduce the number of join

operation and it was six times faster than the HUO-Miner algorithm. Estimated Utility Co-occurrence pruning was developed with the FHM mechanism. The EUCP was used to reduce or eliminate the low-utility extension. Advantages of the algorithm was,

- Reduce the search space
- It was faster compare than HUI-miner

Begum, et al. [12]introduced the pruning strategies for both upper and lower bound. It was produced a robust Dynamic Time Warping (DTW) clustering algorithm which was faster. The similarity search diminishes are used the large data set in the Euclidean distance. It was used in both single and multidimensional domains, such as astronomy, speech physiology, medicine, and entomology. Mukhopadhyay, et al. [13]reviewed the data mining technique for an efficient predictive model for a large amount of data. In multiple conflicting measures of their performances were the main problems in data mining. Multiple objective algorithms were used in many application in the domain of data mining. MOEAs was used in data mining because of their flexibility and it solved the multiple data mining problems. MOEAs was used to solve the two tasks, such as feature selection and classification. Barak and Modarres [14]surveyed the approach for simultaneous prediction of risks in data mining as well as the fundamental data set. In preprocessing the appropriate database was used. The 15 different prediction algorithms were used to predict the risks in data mining. If-Then Rule's gains of the rule-based algorithm were used to analyze the strength and weakness of the tree algorithm. The hybrid feature selection algorithm was used the nine different filter algorithms and function based clustering method. The number of effective features was less compare than a number of effective features on the risk parameter. The limitation of this method was collecting all data and the information was difficult for some real cases.Halder, et al. [15]proposed the subgraph based periodic pattern mining algorithm and it was termed as SPP miner. The algorithm was mostly used in polynomial graph mining. It was included the two pruning method, such as,

- Non- closed pattern pruning
- Non-parsimonious periodicity pruning method

Advantage of the method was,

- Its memory was more efficient compare than other algorithm

To discover the periodic patterns for synthetic datasets and real data set. In this section, discussed various paper and their disadvantages. These drawbacks are overcome by the proposed work.

PROPOSED WORK ;In this section presents the detailed description for the proposed FiDoop using map reduce. In this work, the preprocessing technique is used for remove the error in the data. Next, the information is analyzed and the partitioning is used the map reduce. The threshold is used for the group the limit number of data and the clustering algorithm is used in the confidences. After, that, the frequency item set is used to order the item. Finally, the data are merged and provide the output.

Fig 1. Overall proposed work

A. *Dataset*

An open source data mining library offers various item sets, utility item sets for frequent mining. From the library of item sets, an accident data sets are used in this work. Also, the National Institute of Statistics provided the accident datasets for the region of Flanders (Belgium) during the period of 1991-2000. The datasets collected from the "Analysis form for Traffic Accidents" which occurred with injuries or deadly wounded cases. The traffic data contains several numbers of attributes namely, course and conditions of traffic (type of collisions, road users, injuries, maximum speed, regulation), environmental (time of accident, light and weather conditions), geographical conditions (locations and physical characteristics. The total of 340,184 traffic accidents included in the dataset. The characteristics of dataset described in table I.

Table 1. Data Set Description

Dataset Name	Number of Transactions	Number of distinct items	Average transaction length
Accidents	340184	468	34

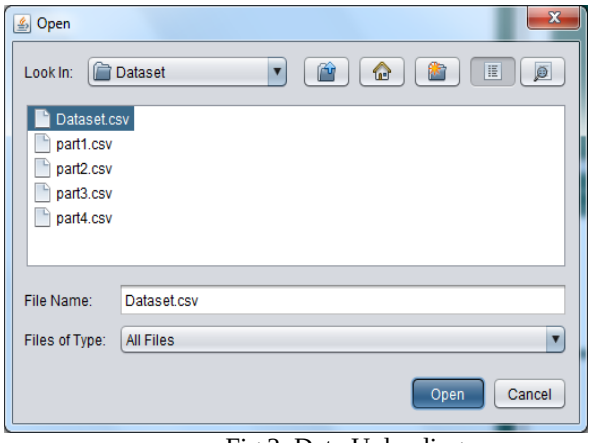

Fig 2. Data Uploading

B. ***Preprocessing***

Association rule mining is used in the preprocessing process. It used to convert the transforming the raw data into an understandable format. The data contain some error means,

it is eliminated by using the algorithm. By applying preprocessing technique it reduces the unwanted data or information. Preprocessing prepares the raw data for further reasons.

Fig 3. Preprocessing

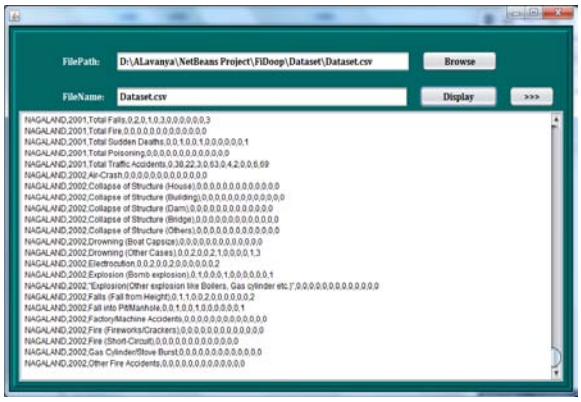

Fig 4.Preprocessing

C. *Partitioning data and threshold*

Partitioning data usedthe map-reduce and it considered as a processing technique. The map reduce algorithm includes two tasks, such as map and reduce. The map is used to convert the convert the dataset from one form to other. Reduce task provide the output from map input and combined the data into a smaller set of data. The sequence of the name map reduce indicates the reduce task is always performed after the map job. The Hadoop in map reduce is used to set the limit of each data.

D. *Clustering algorithm*

The clustering is defined as process of organize the objects into groups. The inputs of clustering algorithm is considered as uncertain database and support and confidence. The output is frequency item set. The algorithm is used to reduce the complexity of the candidate item.

Algorithm 1. Apriori based frequent item set mining
Input: uncertain data base (D), support and confidence, c1
Output frequency item set
Step 1: $\mu = \min(support, Dataset\ (D))$
$C = Generate\ the\ single\ item\ candiates\ of\ D$
K=1, j=0;
Initialize K=1, j=0
While,
Absolute value of$C_k \neq 0$, do
For
each item
$i.\mu = 0$
While $(++j) \leq n\ and\ |C_k| \neq 0, do$
Foreach$I \in C_k$ do
$i.\mu = i.\mu + pr(I \leq t_j)$;
If $I.\mu \geq \mu_m$
$F_k.push\ (I)$;
$C_k$remove(I)
Else if stop
$C_{k+1} = generate\ candiate\ of\ (f_k)$
Return
Frequency item set
End.

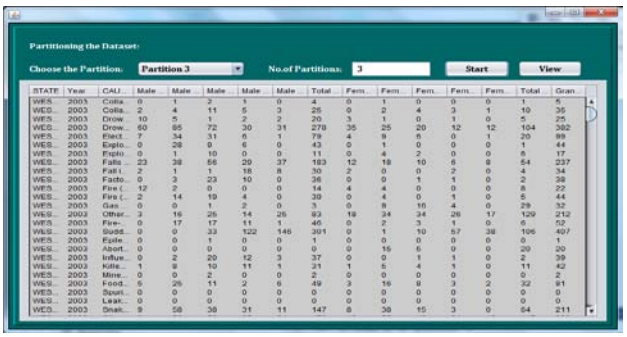

Fig 5. Clustering algorithm

E. *Support and confidence*

Support: The probability of particular transaction with all required items defines the support. Let us consider the two items X and Y. The support declares the probability of occurrence of X and Y in the same transaction. It is defined as,

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

Confidence: The degree of certainty for a given rule refers confidence. Let us consider the confidence$c$. The $c\%$ transactions containing both X and Y specified by the rule$X \Rightarrow Y$. The confidence c is represented as,

$$c = \frac{Number\ of\ transactions\ having\ both\ X\ and\ Y}{Number\ of\ transactions\ containing\ X} \quad (2)$$

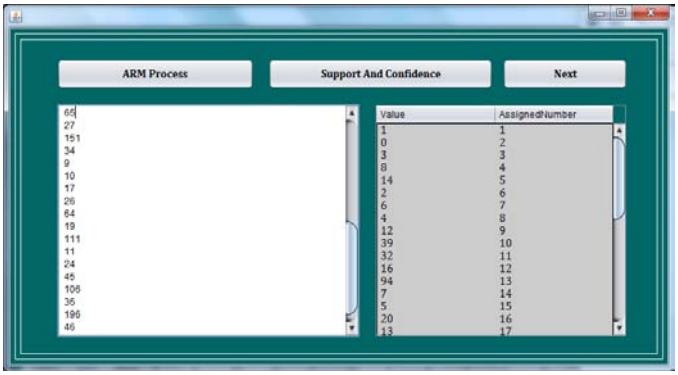

Fig 6.Support and confidence

At last, the documents are ordered and it is display in output.

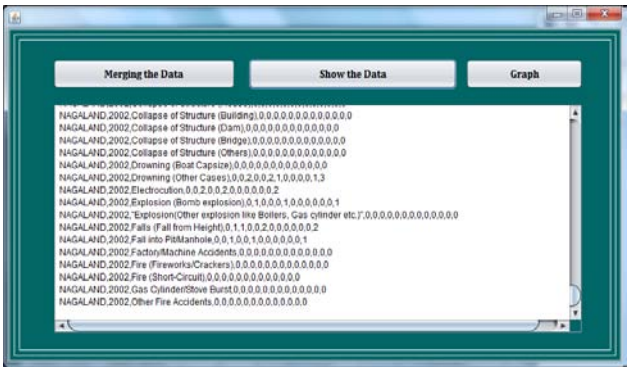

Fig 7. Merge Document

III. PERFORMANCE ANALYSIS

The FiDoop method is simulated in JAVA and it requires WAMP server tool, Eclipse, and Hadoop for IDE and MySQL database. This section presents the results of proposed FiDoop using map reduce. The results are analyzed and evaluated in terms of

- Accuracy
- Memory usage
- Execution time

Moreover, the proposed method is compared with the existing method and the proposed work is implemented in JAVA.

A. Accuracy

Fig 8 shows the comparison of accuracy for both existing and proposed method. The accuracy is increased compared than existing system. Accuracy is defined as the ratio of corrected predictions and t eh total number of predicted values.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ of\ all\ cases\ to\ be\ predicted} = \frac{a+d}{a+b+c+d} \quad (3)$$

$a - true\ postive$
$b - false\ negative$
$c - false\ postive$
$d - truenegative$

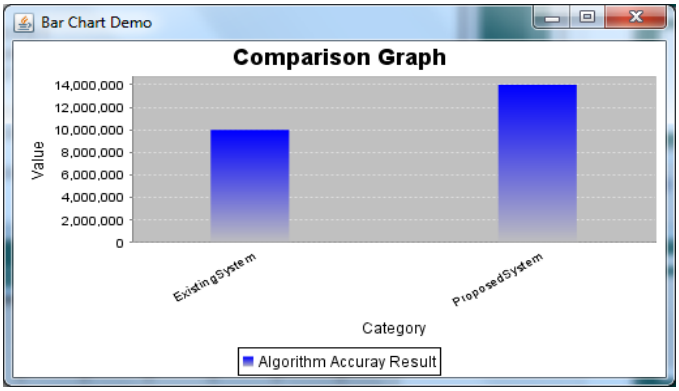

Fig 8.Accuracy

B. Execution time

Fig 9 shows the comparison of execution time for Apriori, FP-Growth and FiDoop. The FiDoop provide the less execution time compared than other methods. If the process are moved during their execution from one memory segment to another memory segment and it provide the delay until their run time. These are generated in hardware and mostly occur in general purpose OS.

$$CPU\ Time = I * CPI * T \quad (4)$$

$I - number\ of\ instructions\ in\ the\ program$
$CPI - Average\ cycle\ per\ instruction$
$T - clock\ cycle\ time$

Fig 9. Execution time

C. Memory usage

Fig 10 shows the memory utilization for Apriori, FP-Growth and FiDoop. The Fi-Doop used the less memory utilization compare than other methods.

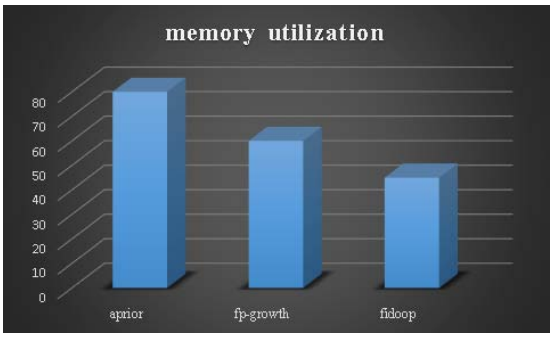

Fig 10. Memory usage

IV. CONCLUSION AND FUTURE WORK

This paper, proposed the FiDoop based parallel mining of frequent item set using map reduce. Initially, the accidental data set is selected and uploaded in the server. After that, the preprocessing technique is used to reduce the error and eliminate the unwanted data. Next, the informations are analyzed and threshold the particular limit of data. These data are classified depends on threshold values and the clustering

algorithm is used to detect the support and confidence values among clusters to discover frequent item. Finally, these frequent item set is used to order the data and the output is merged. The proposed method is used to crop the datafor particular user and it provide the high accuracy, less execution time and less memory utilization. It provides the better result compared than existing of Apriori, and Fp-Growth method. In future, the dimensionality of the nodes are increased in FiDoop.

V.REFERENCES

[1] Y. Xun, J. Zhang, X. Qin, and X. Zhao, "FiDoop-DP: Data Partitioning in Frequent Itemset Mining on Hadoop Clusters," IEEE Transactions on Parallel and Distributed Systems, vol. 28, pp. 101-114, 2017.

[2] Y. Xun, J. Zhang, and X. Qin, "Fidoop: Parallel mining of frequent itemsets using mapreduce," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 46, pp. 313-325, 2016.

[3] X. Yuan, "An improved Apriori algorithm for mining association rules," in AIP Conference Proceedings, 2017, p. 080005.

[4] T. Calders, N. Dexters, J. J. Gillis, and B. Goethals, "Mining frequent itemsets in a stream," Information Systems, vol. 39, pp. 233-255, 2014.

[5] C.-W. Tsai, C.-F. Lai, M.-C. Chiang, and L. T. Yang, "Data mining for Internet of Things: A survey," IEEE Communications Surveys and Tutorials, vol. 16, pp. 77-97, 2014.

[6] J. Yabing, "Research of an improved apriori algorithm in data mining association rules," International Journal of Computer and Communication Engineering, vol. 2, p. 25, 2013.

[7] F. Nori, M. Deypir, and M. H. Sadreddini, "A sliding window based algorithm for frequent closed itemset mining over data streams," Journal of Systems and Software, vol. 86, pp. 615-623, 2013.

[8] G. Lee, U. Yun, and K. H. Ryu, "Sliding window based weighted maximal frequent pattern mining over data streams," Expert Systems with Applications, vol. 41, pp. 694-708, 2014.

[9] G. Pyun, U. Yun, and K. H. Ryu, "Efficient frequent pattern mining based on linear prefix tree," Knowledge-Based Systems, vol. 55, pp. 125-139, 2014.

[10] S. D'Oca and T. Hong, "A data-mining approach to discover patterns of window opening and closing behavior in offices," Building and Environment, vol. 82, pp. 726-739, 2014.

[11] P. Fournier-Viger, C.-W. Wu, S. Zida, and V. S. Tseng, "FHM: Faster high-utility itemset mining using estimated utility co-occurrence pruning," in International Symposium on Methodologies for Intelligent Systems, 2014, pp. 83-92.

[12] N. Begum, L. Ulanova, J. Wang, and E. Keogh, "Accelerating dynamic time warping clustering with a novel admissible pruning strategy," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 49-58.

[13] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. A. C. Coello, "A survey of multiobjective evolutionary algorithms for data mining: Part I," IEEE Transactions on Evolutionary Computation, vol. 18, pp. 4-19, 2014.

[14] S. Barak and M. Modarres, "Developing an approach to evaluate stocks by forecasting effective features with data mining methods," Expert Systems with Applications, vol. 42, pp. 1325-1339, 2015.

[15] S. Halder, M. Samiullah, and Y.-K. Lee, "Supergraph based periodic pattern mining in dynamic social networks," Expert Systems with Applications, vol. 72, pp. 430-442, 2017.