Minimization of Association Rules Using Hybrid Algorithm

Preeti Rani, Nidhi Tyagi

Abstract*: *A Wal-Martsalesman[1] was trying to surge the sales data of the store by combining the commodities together and putting discounts on those products. To discover occasions and additional such goods that can be knotted collected, the sales man examined all sales archives. What he got to know was fascinating. Numerous clients who bought diapers also took beers. The 2goods are clearly distinct thus he found that nurturing kids is exhausting. And to release pain, guardian decided to buy beer. Data Mining, also known as Knowledge Discovery in Databases (KDD) [1], to find irregularities, associations, arrangements, and tendencies to forecast consequences. Apriori algorithm is a standard process in data mining. It is utilised for mining recurrent sets of items and related association rubrics. It is formulated to work on a database comprising of a lot of transactions. It is very vital for operative Market Basket Investigation and this assistance the patrons in buying their substances with more effortlessness which escalates the sales of the markets. While finding goods to be associated together, it is imperative to have some association on which the commodities can be listed together. But this task comes with a drawback because it ends up in so many association rules. In this research work a hybrid method has been proposed to minimize association rules using optimization algorithm Differential Evolution with Apriori Algorithm.

Index Terms*: *Apriori Algorithm, Genetic Algorithm (GA), Differential Evolution (DE), Optimization Algorithm, Data Mining, Association Rules.

I. INTRODUCTION

Mining of frequent set of items [3] is a data-mining technique which is widely used for determining sets of commonly happening items in huge databases. A distinctive instance where this algorithm is utilized is perceive the associations of diverse things in the store. For illustration, to detect which stuffs are accepted frequently collected. This data would allow them to have improved prearrangement plans of their belongings and would also benefit a prodigious deal with effective marketing of goods. The common item set mining (FIM) is also very convenient in numerous request domains such as social networks and supermarket applications to spot tendencies in sales. This applications in significant arenas make common item set mining a zone of active investigation these days.

II. DATA MINING

Data mining is the method of searching irregularities [4], forms and associations within large data sets to predict outcomes. By means of a wide-ranging tools and techniques. In other words data mining is the exercise of constantly probing fat stores of data to notice outlines and trends that go outside modest analysis. Data mining practices cultured mathematical algorithms to slice the data and assess the possibility of upcoming events. Data mining is also identified as Knowledge Discovery in Data (KDD).

A. The Key Assets of Data Mining:

In voluntary detection of forms, forecast of possible consequences, formation of actionable data. Data mining [11] can response questions that cannot be given through guileless query and reporting methods. Data mining is completed by constructing models. A model utilises an algorithm to act on a set of data. The perception of involuntary discovery refers to the implementation of data mining models.

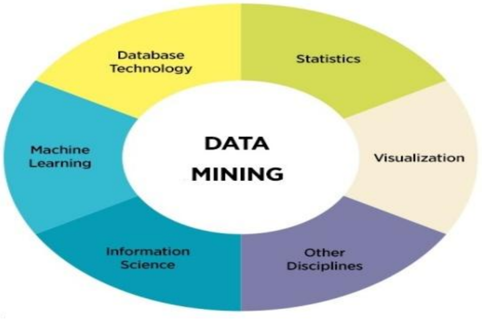


Figure 1: Elements of Data Mining

B. Applications of Data Mining

In business, it gathers data from numerous sources like sales, client buying antiquity, things transportation, ingesting, and services, classifying client behaviours, multidimensional examination of deals, patrons, merchandises, time and region; efficiency of sales movements; customer retention

In the telecommunication business, data mining assistances classify telecom outlines, notice fake events, recover the superiority of services and also make healthier use of possessions [17].

Data mining has also completed important donations to genetic information analysis like genomics, proteomics, functional genomics, and biomedical investigation. It aids in analysis by semantic addition of heterogeneous, dispersed genomic and proteomic databases; connotation and path analysis, picturing tools in genetic information analysis, and more.

Revised Manuscript Received on December 22, 2018.

Preeti Rani First Author name, Department of Computer Science & Engineering, MIET Meerut, India

Second Author name, Department of Computer Science & Engineering, MIET Meerut, India

III. KNOWLEDGE DISCOVERY DATABASE (KDD):

KDD refers [6] to the overall process of discovering useful knowledge from data. It comprises of the evaluation and possibly interpretation of the arrangements to make the choice of what qualifies as knowledge.

A. Steps involved in KDD

(a). Data Cleaning − Basically in this step, the noise and inconsistent data are removed.
(b). Data Selection − basically, in this step, data relevant to the analysis task are retrieved from the database.
(c). Data Transformation −In this step, data is transformed into forms appropriate for mining. Also, by performing summary or aggregation operations.
(d). Data Mining − generally, in this, intelligent methods are applied in order to extract data patterns.
(e). Pattern Evaluation − basically in this step, data patterns are evaluated.
(f). Knowledge Presentation − generally, in this step, knowledge is represented.

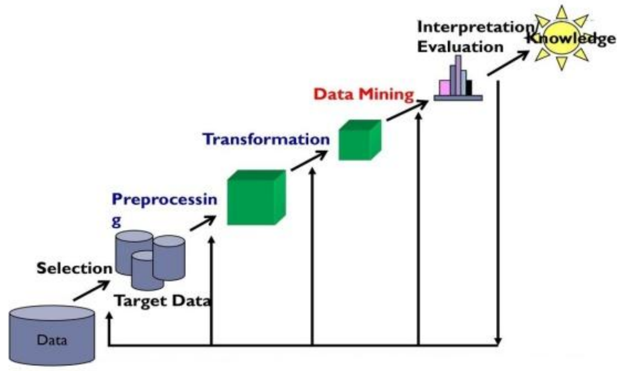


Figure 2: KDD Processing

B. KDD in Data Mining

The procedure of discovering and deducing go outlines from data involves the repeated application of the subsequent tags:

Evolving an understanding of:

- The application sphere
- Relevant previous knowledge
- The objectives of the end-user

Creating a target dataset:

Choosing a data set, or concentrating on a subset of variables, or data samples, on which discovery is to be done.
Data cleaning and preprocessing:

- Elimination of noise or outliers.
- Gathering necessary information to model or account for noise.
- Plans for handling missing data fields.
- Accounting for time sequence information and known variations.

Data reduction and projection:

- Finding useful features to represent the data depending on the goal of the task.
- Using dimensionality reduction methods to reduce the effective number of variables. That is under neat concern or to discover invariant representations for the information.

Picking the data mining algorithm(s):

- Selecting method(s) to be used for searching for patterns in the data.
- Deciding which models and parameters may be appropriate.
- Matching a particular data mining method with the criteria of the KDD process.

Data Mining:

- Searching for outlines of interest in a particular representational form. Such representations as cataloguing rules or trees, regression, bunching, and so forth.
- Interpreting mined patterns.
- Combining discovered knowledge.

IV. OPTIMIZATION ALGORITHMS

Optimization algorithms supports to mineralize (or maximize) an target function which is basically a mathematical function reliant on internal learn able parameters of model which are utilized in computing the objective values from the set of predictors utilized in the model. Genetic Algorithm and Differential Evolution both of these are optimization which assists in minimizing association rules to make the analysis of customer transaction fast, easy and accurate.
Genetic Algorithm in addition to Differential Evolution Algorithm both of these are optimizing method for Apriori algorithm [8]. Apriori algorithm has quite a few association regulations and both of these systems backs the Apriori algorithm by minimizing the association policy of Apriori Algorithm. Knowledge Discovery in Databases (KDD) has been a very charming and fascinating investigation challenge. Its focal point is to draw captivating and determined data from a bulky miscellany of data kept in the transactional databases.

A. Genetic Algorithm (GA)

The core idea of Genetic Algorithm is to imitate the natural choosing and the survival of the fittest. In Genetic Algorithm, the solutions are shown as chromosomes. The chromosomes are evaluated for fitness values and they are graded from best to worst based on fitness value. The procedure to produce new solutions in GA is to replicate the natural selection of living organisms [16], and this process is accomplished through repeated applications of three genetic operators: selection, crossover, and mutation.

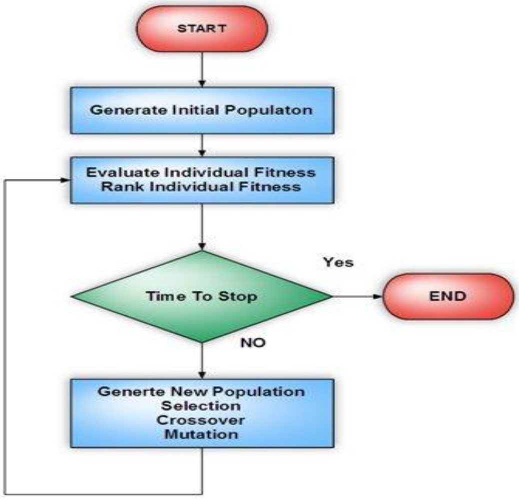


Figure 3: Genetic Algorithm Flowchart

B. Differential Evolution (DE)

There are 3 main processes in all evolutionary algorithms. The first process is the starting process where the basic population is randomly produced [14] according to some solution representation. Every individual represents a solution. If an indirect representation is used, everyone must first be decoded into a solution. Every solution in the population is then evaluated for fitness value in the 2nd phase. The fitness values can be used to define the average population fitness or to rank the individual solution. The 3rd process is the creation of a novice population by perturbation of solutions in the existing population.

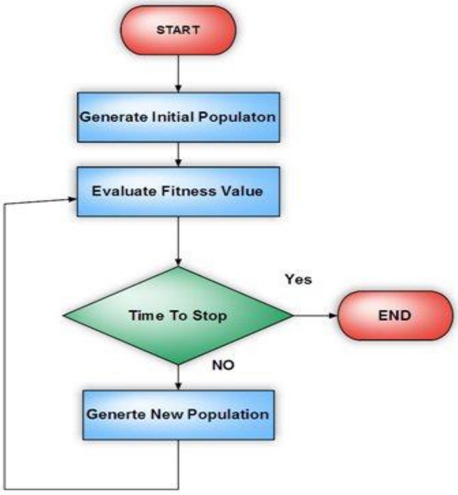


Figure 4: Different Evolution Flowchart

C. Reduction of Items to Find Best Associations

The Apriori theory [9] can decrease the number of item sets that are needed to inspect. Put simply, the Apriori principle states that if an item set is infrequent, then all its supersets must also be infrequent, this means that if {beer} was initiate to be infrequent, it can be anticipated {beer, pizza} to be uniformly or even more infrequent. So in merge the list of popular item sets, it is needed not to deem {beer, pizza}, nor any other item set formation that contains beer.

Finding item sets with high assistance, Using the Apriori principle, the number of item sets that have to be scrutinized can be pruned, and the list of frequent item sets can be obtained in below steps:

Step 0. Start with item sets having just one item, such as {apple} and {pear}.

Step 1. Decide the support for item sets. Keep the item sets that gather your minimum support threshold, and remove item sets that don’t.

Step 2. Utilizing the item sets you have kept in Step 1, produce all the possible item set configurations.

Step 3. Reiterate Steps 1 & Step 2 till there are no more novice item sets.

{apple} was decide to have low support, hence it was removed and all other item set configurations that contain apple need not be considered. This abridged the number of item sets to consider by more than half.

Note that the support threshold that you pick in Step 1 could be foundation on formal analysis or past experience. If you find out that sales of items beyond a certain proportion tend to have a noteworthy impact on your profits, you might consider using that amount as your support threshold.

V. PROPOSED METHODOLOGY

In this proposed work we’ve abated association rules considering Differential Evolution. Since the detection of interesting association relations among huge amounts of sales transactions is now vibrant for making suitable business verdicts. There are presently a diversity of algorithms to notice association rules. Some of these algorithms depend on the use of minimum support to weed out the uninteresting rules.

Here we implement two hybrid algorithms apriori with genetic algorithm and apriori with differential to minimize the association rules. The steps of algorithms are below-

A. Apriori with Genetic Algorithm (GA)

GA is based on the principle of natural selection and development. A GA is a process heuristic that imitate the procedure of natural evolution. This process is used to get purposeful solutions for optimization and search related problems. GA is a type of evolutionary algorithms (EA), which come up in each step to give solutions to optimization problems by deploying various techniques motivated by natural evolution like mutation, crossover, inheritance and selection.

Algorithm :

1. Begin
2. Load dataset in the memory.
3. Deploy Apriori Algorithm on it to frequent product groups.
4. Suppose F is the frequent item-set set gained by Apriori Algorithm.
5. Set O =Φ where O is the o/p set having all discovered association rules.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

6. Apply some terminating rules on Genetic Algorithm.
7. Put each item set of F in some encoding policy.
8. Then, choose members and apply GA on them to produce association rules.
9. No, find the fitness function of every rule A→ B.
10. If cost of fitness function come across the criteria of choice then
11. Set O= O {A→ B}.
12. If the required number of generations is not completed, then move to step 3.
13. END.

B. Apriori with Differential Algorithm (DE)

Algorithm :

1. **Begin**
2. Load an item-set to the memory.
3. Deploy Apriori Algorithm on it to frequent item-sets. Let F is the frequent item-set set achieved by Apriori Algorithm.
4. Set O =Φ where O is the o/p set having all discovered association policies.
5. Apply some terminating regulations on Differential Evolution Algorithm.
6. Show each item set of F in some encoding scheme.
7. Then, chosen members and apply DE Algorithm on them to produce association rules.
8. Now, calculate the fitness function of each policy A→ B.
9. If value of fitness function meet the level of selection then
10. Set O= O U {A→ B}.
11. If the required number of generations is not finished, then move to step 3.
12. END.

C. Association Rule Mining (ARM)

Association rules are intended to discover strong rules from databases with the help of a variety of measures of interestingness and for uncovering regularities and sturdy relation among items in huge transactional data . ARM focuses on identifying appealing correlations, frequent occurring pattern, associations or informal structures between sets of items in the commercial databases or any other data repositories. Apart from this ARM assists in separating correlations amid products belonging to any customer conducting business in some market -basket DB can be effectively discovered using ARM.

Regulations in Association Rule Mining algorithms are usually in the form: X→ Y.

IF the value of the forecasting attributes is true, THEN value is predicted for goal attributes.

Both X, Y are frequent item-sets in some DB and X ∩Y= ∅. The rule X→Y can be explained as "if some item-set X happens in a transaction, then some extra item set B will also befall in the same transaction". For instance, suppose in some DBs 35% of total transactions comprise of both bread and sauce and 75% of all transactions include bread. An ARM system will create the rule bread → sauce with 35% support and 75% confidence. Rule support and rule confidence are 2 very vital virtue factors of rule interestingness.

D. Performance Evaluating Parameters

These are the explicit events used to find the association rules (in market-basket analysis) amid various items. For instance consider 2 items X and Y then,

Support : This gives the frequency of the item in the dataset. The support of an item set Y, shown as (Y), can be said as the percentage of transactions in the dataset, having item set Y.

Y = σ (AUB) / σ (N), Where (N) = transactions in the dataset (AUB) = number of transactions that has both A and B

If you deem a basket containing 10 items (5-apples, 3-eggs, 2-pens) then support of any precise item say apple can be 5 as mentioned. Similarly precise value can be intended by the proportion of number of occurrences to the total number of items in the basket (i.e., support (apples) = 5/8).

Confidence : : This clarifies how likely, Y is likely to be bought when X is bought. This defines association between two items.

Confidence is additional degree to evaluate the precision of association rule. It calculates the conditional probability.

C= σ (AUB) / σ (A), Where σ (A) = number of transactions comprising A.

A very operative association is considered a mid A and B if the confidence value is greater.

For instance when a person buys milk is more likely to buy bread as well or vice versa. This is restrained by the proportion of relations with item X, in which item Y also appears. It is expressed as {X→Y}. Premeditated by the proportion of number of transactions in which both (X & Y) occurs to support of the item X.

Comprehensibility- If produced rule has a huge number of attributes with it, then that regulation will be considered as so tough grasp. The discovered rules should be very easy and comprehensible to the user, so that the user can use them effectively. So the Comprehensibility factor is much needed for making rules a bit easy to grasp. Comprehensibility of a certain association rule can be defined as:

Comp =log (1+ B) / log (1+ AUB)

Where B and |AUB| are the number of attributes integrated in the coming part of the rule and total rule respectively.

Interestingness

Interestingness of certain rule, represented as Interestingness X→Y, is used to quantify how much a certain rule is "eye-opener" for the users. Since finding some concealed data is the core point of data mining, so it should reveal those regulations that have relatively less occurrence in the DB. Interestingness of a rule can be formulated as:

Interestingness X→Y

= [(Sup(XUY)/Sup(X)) x ((Sup(XUY)/Sup(Y)) x (1-(Sup(XUY)/ σ(N))]

Where σ (N) indicates the total number of transactions in the database.

As stated, Association Rule Mining is reviewed as Multi-objective difficult no Single Objective one. So, its fitness function can be formulated as:

F= ((P × Support) + (Q ×Con.) + (R × Comprehensibility) + (S × Interestingness)((P+Q+R+S))

Where P, Q, Rand S are user-defined weights.

Since ascertain frequent item sets is of high computational complexity so, dig out association rules can be abridged to finding frequent item sets. In this project work the weight values of P= 4, Q=3, R=2 and S=1 are taken based on the relative significance of the 4quality measures support, confidence, comprehensibility and interestingness. It should also be noted that the range of fitness values should be in [0…1].

VI. EXPERIMENTAL RESULTS

By analyzing the below experimental results, it is obvious to say that Apriori Different Evolution Algorithm is better than Genetic Apriori Algorithm. Multiple datasets have been used for the research process and to find the results.

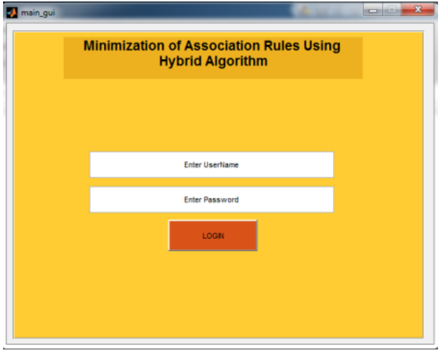


Figure 5 : GUI in Command Window.

As shown in above figure, the GUI of a project is created to present the project along with Login credentials.

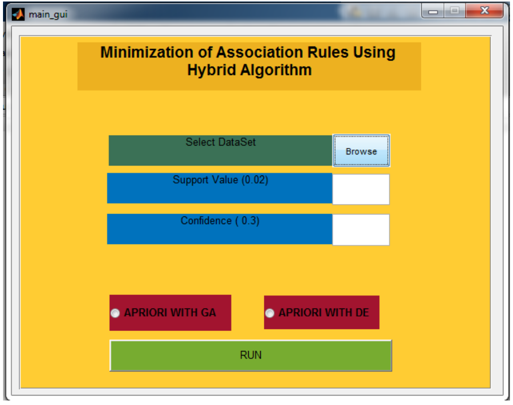
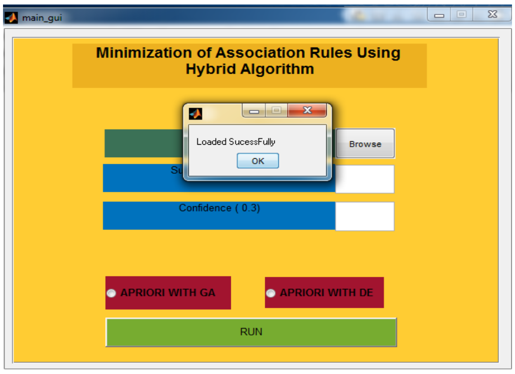


Figure 6: Main Screen after Login

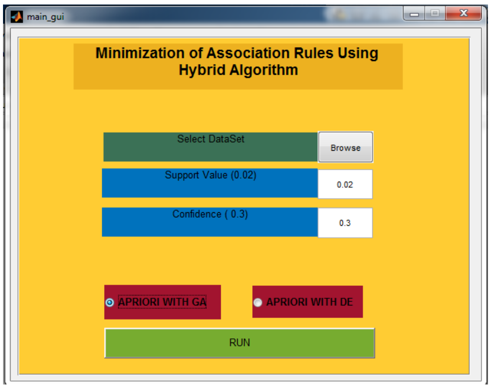
As shown in figure 6, this is the window for browsing the dataset which is used in our project, after selecting the dataset, both algorithm perform the operations on it and show the results.



Loaded SucessFully

Figure 7: Screen after Loading Dataset

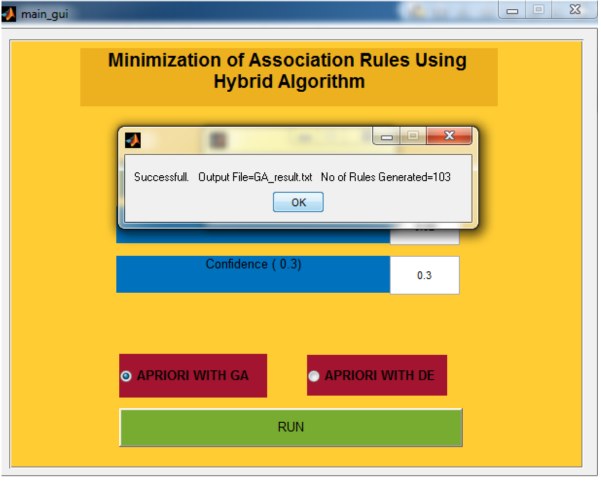
As shown in figure 7, this window is shown our dataset is loaded successfully. Dataset is loaded when it is clicked on the browse button, here any dataset can be selected to run with both algorithms to compare our results in the form of minimization of association rules.



Support Value (0.02): 0.02
Confidence (0.3): 0.3
● APRIORI WITH GA

Figure 9: Select the Algorithm

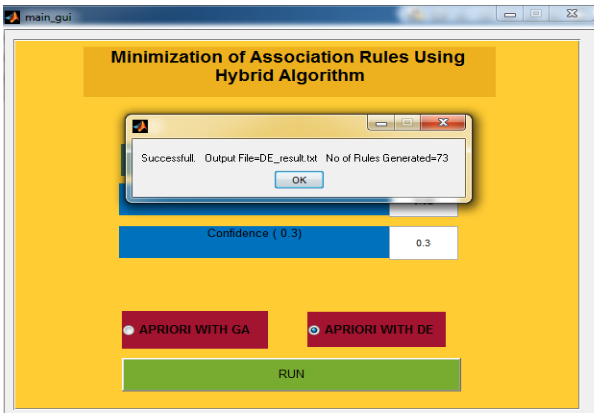
As shown in figure 8, in this window, any of the 2 algorithms can be chosen, click on the run button to execute the algorithm with respective dataset and parameter values to compare with another one algorithm in the form of minimization of association rules.



Successfull. Output File=GA_result.txt No of Rules Generated=103

Figure 9: Results after running dataset, Apriori with GA.

After running dataset with the Genetic Algorithm the total association rules are 103.



Successfull. Output File=DE_result.txt No of Rules Generated=73

Figure 10: Results after running dataset, Apriori with DE.

But when same dataset is executed with the proposed Hybrid DE Algorithm then total association rules reduced to 73.

VII. RESULT

As it can be seen in the Table that our proposed technique is efficient, since it has reduced and has lesser rules as compared to Apriori with GA. As we can see that our proposed method Hybrid priori with DE has only73 association rules and Apriori with GA has 103 association rules, thus it proves proposed technique is far better.

Table 1: Result Analysis

Algorithm	Support Value	Confidence Value	No. Of Rules
Apriori with GA	0.02	0.3	103
Apriori with DE	0.02	0.3	73

VIII. CONCLUSION & FUTURE WORK

Partitioning of data is an important in data mining and its objective for formation of dominant group is to analyze the input data and develop an accurate model for each group using features present in the data. And in this research work it has been proved with this proposed hybrid method using Apriori algorithm with Difference Evolution.

Though this research work has been successful in addressing the problem of achieving higher prediction rate, less execution time and lower memory consumption with the help of proposed modules, this research work has not taken the following aspects into account:
•Setting up multiple minimum support values for the items.
•Generation of only user interested association rules.
•Setting up of different minimum support values of various items present in the database.

REFERENCES

1. U. Fayyad, G. Piatesky-Shapiro & P. Smyth, "From Data Mining to Knowledge Discovery in Databases",AI Magazine, 17(3):37-54, Fall 1996.
2. Q. Yiang and X. Wu,"10 Challenging Problems in Data Mining Research", International Journal of Information Technology & Decision Making, Vol. 5, No. 4, 2006, 597-604.
3. L. Page, S. Brin, R. Motwani, T. Winograd," The Pagerank Citation Ranking: Bringing Order to the Web, ", Technical Report, Stanford University, 1999.
4. M. E. J. Newman,"The Structure and Function of Complex Networks ",SIAM Review, 2003, 45, 167-256.
5. L. Getoor, "Link Mining: A New Data Mining Challenge "SIGKDD Explorations, 2003, 5(1), 84-89.
6. https://www.slideshare.net/INSOFE/apriori-algorithm-36054672.
7. https://en.wikipedia.org/wiki/Association_rule_learning.
8. Amin. A. and S. Fisher. "A document skew detection method using the Hough transforms. Pattern Analysis & Applications", 3(3):243–253, 2000.
9. https://www.geeksforgeeks.org/apriori-algorithm/.
10. Voratas Kachitvichyanukul, "Comparison of Three Evolutionary Algorithms: GA, PSO and DE" IEMS August 18, 2012
11. Boroujeni S A, "Learning with Labele and Unlabeled Data", Master's Thesis, Amirkabir University of Technology (Tehran Polytechnic), 2000.
12. N. Chawla, G. Karakoulas, "Learning from Labeled and Unlabeled Data: An Empirical Study across Techniques and Domains ", Journal of Artificial Intelligence Research, 23:331-366, 2005.
13. Kevian Borna, Vahid Haji Hashemi, "An Improved genetic Algorithm with A Local Optimization Strategy and Extra Mutation Level For Solving Travelling Salesman Problem", IJCSEIT, Vol. 4, No.4, August 2014.
14. Swagatam Das, Sankha Subhra Mullick et al., "Recent Advances in Differential Evolution – An Updated Survey", Swarm and Evolutionary Computation · February 2016.
15. Yu-Xiang Lei, Jin Gou et al., "Improved Differential Evolution With a Modified Orthogonal Learning Strategy" IEEE Access May 25, 2017.
16. Bingyan Mao, Zhijiang Xie et al., "A Hybrid Strategy of Differential Evolution and Modified Particle Swarm Optimization for Numerical Solution of a Parallel Manipulator", Article ID 9815469, Volume 2018.
17. Nidhi Tyagi, R.P.Agarwal and Rahul Rishi "Context based Web Indexing for Storage of Relevant Web Pages", International Journal of Computer Applications, Volume 40-No.3, 2012.
18. Shailza Chaudhary, Pardeep Kumar, Abhilasha Sharma, Ravideep Singh, "Lexicographic Logical Multi-Hashing for Frequent Itemset Mining", International Conference on Computing, Communication and Automation (ICCCA2015)
19. Feng Gui, Yunlong Ma, Feng Zhang, Min Liu, Fei Li, Weiming Shen, Hua Bai, "A Distributed Frequent Itemset Mining Algorithm Based on Spark", Proceedings of the 2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD).