Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

An improved hidden behavioral pattern mining approach to enhance the performance of recommendation system in a big data environment

P. Shanmuga Sundari [a,*], M. Subaji [b]

[a] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632014, India
[b] IIIP, Vellore Institute of Technogy, Vellore 632014, India

ARTICLE INFO

Article history:
Received 8 July 2020
Revised 11 September 2020
Accepted 18 September 2020
Available online xxxx

Keywords:
Hidden Behavioral analysis
Big data
Fp-Growth
Association rule mining
Two-level clustering

ABSTRACT

The proposed work aims to solve data sparsity problem in the recommendation system. It handles two-level pre-processing techniques to reduce the data size at the item level. Additional resources like items genre, tag, and time are added to learn and analyse the behaviour of the user preferences in-depth. The advantage of the proposed method is to recommend the item, based on user interest pattern and avoid recommending the outdated items. User information are grouped based on similar item genre and tag feature. This effectively handle overlapping conditions that exist on item's genre, as it has more than one genre at initial level. Further, based on time, it analyses the user non-static interest. Overall it reduces the dimensions which is an initial way to prepare data, to analyse hidden pattern. To enhance the performance, the proposed method utilized Apache's spark Mllib FP-Growth and association rule mining approach in a distributed environment. To reduce the computation cost of constructing tree in FP-Growth, the candidate data set is stored in matrix form. The experiments were conducted using MovieLens data set. The observed results shows that the proposed method achieves 4% increase in accuracy when compared to earlier methods.

1. Introduction

Amplified growth in web technology leads to over loaded information. Finding the relevant and refined information from this overloaded environment is a big challenge for online user. Recommendation systems play a vital role to suggest such information to the user (Kumar and Thakur, 2018). A wide range of web applications like online news, e-commerce, online music, Netflix, YouTube, Facebook and scientific research, utilize the recommendation systems for better decision making process. Content based filtering method and Collaborative filtering method handle the problems arise in recommendation system (Lu et al., 2015). Some researcher combine these methods to form a hybrid filtering method to achieve better recommendations (Manogaran et al., 2018). The content based method works purely on the basis of similarity in the features of the items and it requires detailed information about the items. Based on similarity, it ranks the items and suggest the topN items to the users (Isinkaye et al., 2015). The collaborative filtering method works only on star rating, which is the user preference score. Rating or user preference is a numerical scale value which is given by the user towards the items that ranges from 1 to 5. If the user is less satisfied or not interested in a particular item, then they rate it as "one". Similarly if the rating score with "five", then it is an indication that the user is more interested or satisfied with the item (Bobadilla et al., 2013). The collaborative filtering method is further classified into memory based and model based (Bobadilla et al., 2013). Memory based method is also called as nearest neighbour method. It calculates similarity distance between the user and the item to provide recommendation. Model based recommendation system is used widely in machine learning algorithm and mathematical models. Due to increase in users and items sparsity in rating matrix, the collaborative filtering method delivers poor recommendation. To overcome this problem it is necessary to incorporate other features to learn implicit user preference. This provides an additional information and evidence about the user preference towards the items. Data mining techniques like clustering, classification and SVD are predominantly used to solve sparsity issue in recommendation system. These methods suffer with high computation cost to train the model for prediction and it delivers less accuracy due to data sparsity problem. Analysing hidden correlation among the user interested items helps to understand the user interest behaviour or pattern. This interest pattern behaviour helps to enhance

* Corresponding author.
E-mail address: pshanmuga.sundari2014@vit.ac.in (P.S. Sundari).

https://doi.org/10.1016/j.jksuci.2020.09.010
1319-1578/

recommendation accuracy. Mining interest pattern in the form of association rules is a very significant technique in the data mining domain. It involves, finding frequent item sets and condition relational rules among the frequent items. These rules helps to discover hidden pattern which advances the excellence of many commercial decision making process like market basket analysis, catalogue design and cross-marketing. This study proposed a novel approach based on hidden behavioural analysis on user preferred items and feature associated with user profile to handle the multi labelled item as well as the items correlations. The hidden behavioural analysis is examined using pattern mining approach. Pattern mining is a basic data mining technique to discover hidden correlations in the data set and detects the frequent item sets. The rules provide the if-then pattern among the frequent item set. The main objective of the proposed work is to handle two-level pre-processing techniques to reduce the data size at the item level. To obtain the improved user profile, additional resources like items genre, tag and time are used to categorize the items for behavioural analysis. At initial level, the items are grouped based on similar item genre, tag feature and other user information. It effectively handles overlapping conditions that exist on movie's genre as it has more than one genre. Based on time, it analyses the recent user interest as their interest are not static. It reduces the dimensions at item level using the clustering technique which is an initial way to prepare the data for association rule mining. Based on rules generated from the association rule mining algorithm, it classifies and predict user preference for better recommendations

The major contribution of the research study is summarized as follows.

- The novelty of the proposed research study is to solve data sparsity problem in collaborative filtering method. Initially the item level data reduction is handled with the help of clustering technique
- User recent preference is calculated with additional features like time and the items are categorized according to their interest.
- Parallel and distributed FP-Growth algorithm are used to enhance the performance of the recommendation system
- Instead of storing transactional database in FP-Growth, the frequent items are stored in the form of matrix, which reduce the memory consumption while constructing a FP-Tree. It reduce the number of scans in candidate data sets.

The significance of selecting the market basket analysis is to discover hidden preference patterns which can be useful in market basket analysis, decision making system, recommendation system, medical treatment and more. Adopting this pattern mining approach in recommendation system can provide specific recommendations or promotions to induce the customer to opt the items. The pattern mining approach learn and examine most important hidden association or correlation of items from the transactional database. From the discovered pattern, the user implicit preferences are well identified. Association rule mining is similar to item based recommendation system. When the product or items occur together or frequently in a basket, the user preference level is high. Association rule mining predicts the hidden correlation based on frequently preferred items from the basket of recent interest. Traditional collaborative filtering method analyse the user preference based on latent feature of preferred items from the past. Analysing the hidden pattern that exists in the user preference matrix enhance the recommendation accuracy. The same pattern is analysed using association rule mining method.

The remaining paper is organized as Section 2, Section 3, Section 4, Section 5 and Section 7.

2. Literature review

Many online applications that includes but not limited to online business, online news, online Music and health care system, utilize the recommendation system to personalize the service and product, based on user requirements and interest. Increase in users need and demand for quality products has become a big challenge for the recommendation system. Series of Algorithms and methods are being developed by researchers to enhance the efficiency and to solve the problems in the recommendation system. Analysing hidden knowledge can be achieved using frequent pattern mining (FPM), sequential pattern mining (SPM) and high utility item set mining (HUIM). Frequent pattern mining approach plays a significant role in extracting hidden data pattern from the data (Luna et al., 2019). Advancements in technology gave birth to many algorithms like sequential pattern mining, multi-threaded pattern mining, distributed and parallel pattern mining approach for better performance.

The study stated that ''Association rules are one of the major techniques of data mining (Solanki and Patel, 2015). It finds frequent patterns (FP), associations, correlations or informal structures among sets of items or objects in transactional databases". It supports algorithms like Apriori, FP-Tree and Fuzzy FP-Tree. Recently Apache's Spark introduced a distributed FP-Growth algorithm for finding frequent item set in the transactional database (Kumar and Mohbey, xxxx). FP-Growth reduces the numbers scan in the transactional database when compared to Apriori algorithm. This algorithm omits the data that are not frequent. Hybrid approach by combining clustering and association rule mining techniques to solve the data sparsity problem (Najafabadi et al., 2017). Based on the user's profile and item's profile the dimension space was reduced at item level. Association rule mining analyse the interested pattern for user preferences. This method adopts the secondary data sources like tags to enhance the preference accuracy level. Association rule mining eliminates the dependencies so that it increase the precision values. The rule generation is high when it handle large data set and it increase the complexity of the prediction.

LAC, is a new Java Library for Associative Classification that includes numerous classification algorithms. It provides the open source framework to evaluate the classification problem (Padillo et al., 2020). Pattern mining approach that works well in small scale data for twitter's hashtag recommendation. It involves two stage process (Belhadi et al., 2020a). In the first stage, the collected twitter data is organized and transformed into transactional data base to discover the hidden pattern. In the second stage, most relevant hashtags are identified and recommended. ''Decomposition Transaction for Distributed Pattern Mining (DT-DPM)" is a distributed pattern mining approach to discover hidden correlation among the transactional database for big data processing with different architecture (Belhadi et al., 2020b). This DT-DPM method initially decompose the transactional data base into different clusters. The clusters are then evaluated with different architecture like single CPU, multiple CPUs and MapReduce framework. DT-DPM achieves better result when compared to Different Pattern Mining approaches namely FIM, WIM, UIM, HUIM and SPM. A clustering based pattern mining approach which discovers the correlation between transactions in transaction database. Highly correlated transactions are grouped using k-means algorithm. Pattern mining approach is applied to find relevant hidden patterns (Djenouri et al., 2019). A hybrid frequent itemset mining (HFIM) using Apriori algorithm is a big data approach which reduces the number of scan in candidate generation. It achieves good scalability as the data are stored in Hadoop distributed file system (Sethi and Ramesh, 2017). Adopting association rule mining in recommendation

system provides more accurate results. In association rule mining approach the support and confidence metrics are the two important factor to affect the accuracy. Its computation cost becomes high while it generates huge number of rules. "Multi-objective Particle Swarm Optimization (MOPSO)" improves the quality of the recommendation system by treating support and confidence as different objectives (Tyagi and Bharadwaj, 2013). The computation cost is reduced as its mines only specific items which are associated with the rules. Finding frequent pattern mining for large data set using sequential data processing is complex as it requires long execution time and high memory consumption. Parallel frequent pattern mining which enhance performance as it distribute the data and processing in a multiple systems (Miao et al., 2019). A parallel improved Apriori algorithm for frequent pattern mining approach (Yang et al., 2015) which introduce new data structure called "Key-value" pair which reduce the number of scans in traditional Apriori algorithm. It occupies less memory, when compared to transactional database. "Apriori-Growth" which is an efficient frequent pattern mining approach by combining Apriori algorithm with FP-tree (Wu et al., 2008). This eventually reduces the computation cost. MR-Apriori is a map reduce based algorithm which solves scalability and efficiency of association rule mining. Finding association rules based on MapReduce framework enhance the efficiency (Lin, 2014).

3. Proposed method

The proposed method identifies the user preferences based on available data sources such as 1.User profile, 2. Item profile, 3. Tags, and 4. Rating. It examines the hidden pattern in the user and item matrix that integrates the user behavior of preferred movie features. By adopting association rule mining it reduces the data sparsity problem. The proposed method is carried out in three stages, as shown in Fig. 1. In the initial stage of pre-processing, the incomplete data is removed and the items are aggregated to reduce the dimension of the item space. Defining the number of cluster size is a challenging task in this stage. The item's feature associated with genre helps to categorize the movie. Hence, the presented method initially considers the number genre as the cluster size and the same is validated with DB index (Maulik and Bandyopadhyay, 2002). The second stage is to discover hidden patterns involved in user preference behavior or interested item. Fp-Growth finds the frequent item set from the transactional database which are essential to generate frequent item sets that are strongly correlated to each other. The output of frequent pattern mining is the input to the association rule mining that build strong if-then association rule. It predicts the unknown user's preferences based on the rule generated by the association rule and analyse the matching behavior pattern of the users and their watched movies. To enhance the accuracy, the strongest rules were selected by applying pruning. The final stage is the recommendation, where the associative classifier is built with help of training data set. Antecedent represent the features and consequent represent the class label. The model is tested with test data and generate topN recommendations.

3.1. Pre-processing

Pre-processing is the initial stage to organize the data for prediction. Based on movie's genre, the data is categorized. As each movie has more than one genre, it leads to overlapping condition. So it is impossible to categorize the movies into a single genre. For example, the movie named "Jumanji" falls into the following genre as Adventure, Children and Fantasy. So categorizing the movie into the proper label is a challenge. To handle this overlapping condition, suggested time-weighted links for community detection, based on user group. Adopting weight over feature helps to solve the overlapping condition (Moradi et al., 2016). So the proposed method adopted nearest neighbour based method. In this method similar user preferred items are grouped.

3.1.1. Bisecting KMeans clustering

Bisecting KMeans clustering is a scalable type of hierarchical clustering algorithm.Spark mllib supports this clustering approach which is used to group the items. The cluster is formed based on item's similarity, where it calculates the similarity between item's genre, tag and time. The cluster size is an important parameter and it is leant by experiment and validated with DB index as given in Fig. 2. The lowest DB index value is the indication of optimal cluster size. From the Fig. 2 the cluster size is fixed as ten. Cluster is named according to the highest frequency of genre existing within the cluster. Each cluster is again grouped based on the time feature that shows the recent interest of the user. User is categorized into three categories as Recent, Medium and Old based on the time feature within the cluster. To achieve this categorization the item's life span is calculated based on difference between the starting year and current year. Threshold is fixed based on average of time difference. If the item's time differences is less than the average, then it is classified as "Old". If the item's time is greater than the average time then is classified as "Recent". When the condition is neutral, the data items is categorized as "Medium". User preference may change over time and the aggregated items are ranked depending on the time. "Recent" preferred product will get a higher ranking. As rating is represented in the numerical form in the MovieLens data, it is converted into categorical data. Based on preference level the item is categorized as "High", "Low" and "Middle" using Eq. (1). These categorizations helps to enrich the user profile and reduced item space to enhance the association rule mining process. The grouping and categorizing data help to simplify the association rule mining process. After pre-processing, an example of enriched user profile is shown in Table 1.

$$Preference\ Level = \frac{Rating\ of\ item\ with\ in\ the\ cluster}{Maximum\ rating\ from\ a\ user\ profile} \quad (1)$$

3.2. Generate association rule

The next phase after pre-processing is to predict the hidden preference of the user. There are two stages to find hidden pattern.

1. Finding frequent item set
2. Generate association rule

3.2.1. Frequent item sets

Let $I = \{I_1, I_2, I_3, \ldots I_n\}$ be the set of items.

D is the transactional database which is represented as $D = \{T_1, T_2, T_3 \ldots\ldots T_m\}$; where $T_j (j \in (1 \ldots m))$ is a user preference (which is same as transaction in market basket analysis) which includes set of items in I. The support is a measure which occurrence of pattern P in D. P is a frequent item set if P's support count is not less the threshold value of minimum support.

3.2.2. Enhanced FP-Growth algorithm

FP-Growth algorithm is an efficient and scalable method. It works on the principle of divide and conquer method. It scans the database twice. The frequent item set are examined and stored in descending order during the first scan. It construct the FP-tree for storing data for frequent patterns in the second scan. The proposed enhanced FP-Growth algorithm converts the frequent item

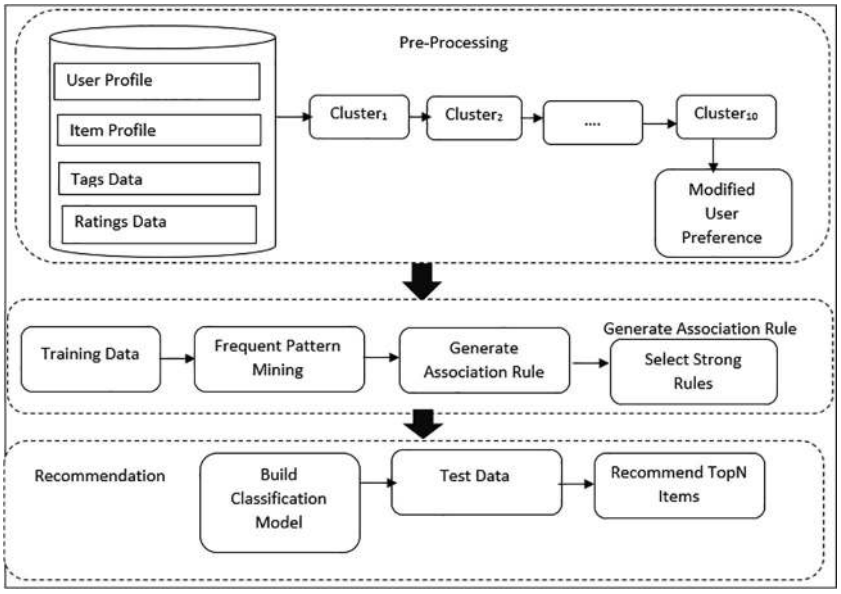


Fig. 1. Proposed Architecture.

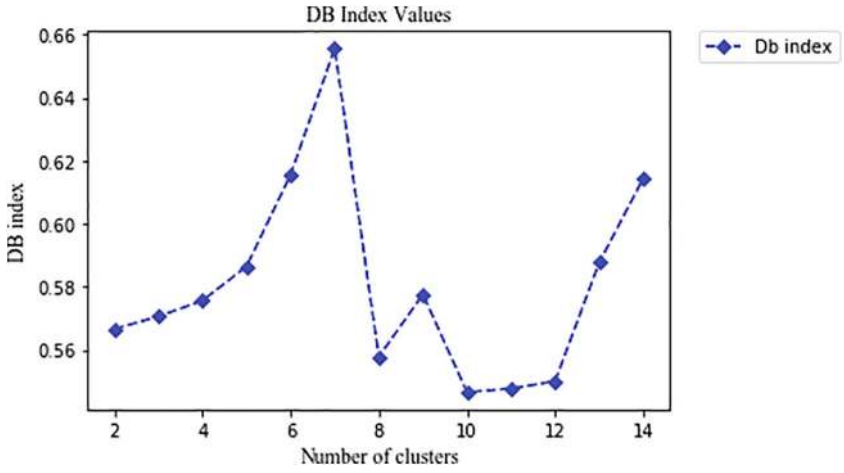


Fig. 2. DB Index.

Table 1
Generated User Profile(After Pre-Processing).

User	Items	Ratings	Group	Preference Level	Ratings to Categorical data
1	100	3	Adventure/Recent	0.6	Middle
1	108	5	Romance/Medium	1	High
1	120	5	Fantasy/Old	1	High
2	160	3	Children/Recent	1	Middle
2	1901	1	Animation/Old	0.33	Low

sets into matrix form. For example consider the following transaction Table 2.

3.2.3. Frequent itemset

Frequent pattern mining scans the entire data to find the frequent item sets which satisfies the minimum support count threshold. Choosing optimal threshold value enhance the accuracy of the association rules which is explained in Section 4.2. The enhanced FP-Growth method is explained with the help of small transaction Table 2. In the database there are collection of transactions which is represented as the $T_id = \{t1, t2, \ldots t5\}$ where T_id is the transaction identifier. The minimum support count threshold is fixed as 3 for this example dataset. The frequent item sets are

Table 2
Sample dataset.

Tid	Original Item sets
*T*1	bread, butter, jam,biscuits, beer, donuts,Sauce, milk
*T*2	butter, eggs, jam, bread, soda, sauce, oil
*T*3	Eggs, bread, diaper, coke,oil
*T*4	eggs, jam, cookies, soap, milk
*T*5	butter, bread, jam, donuts, sauce, nuts

arranged in descending order with their support count. Hence, the frequent item sets are stored in a list called Frequent_list = [(bread:4), (jam:4), (butter:3), (eggs:3), (sauce:3), (milk:3)].

The list consists of item and its support count. The data set is updated based on Frequent_list. Hence other items that do not satisfy the support count threshold are removed from the data. To enhance the efficiency of the proposed method, the updated data set is converted in the form of matrix. The data set in the above Table 3 contains fives transactions and the Frequent_list contains six elements. Hence the matrix is constructed with 5×6 dimensions and represented as *M*. Initially the null matrix *M* with $q \times r$ is created. Where *'q'* is denoted as number of transactions and *'r'* is represented as number of items in the Frequent_list. The matrix *M* get entries from Frequent_list. The proposed method converts the database into matrix form instead of updating the transactional database. For example transaction *T1* is converted in matrix form as it compares the items in the Frequent_list. From the Frequent_list, each item is compared with transaction table, if the entry is present in the list then the corresponding row and column is filled with '1' otherwise it is represented as '0'. The same can be repeated for all the transaction in the database D. The matrix form in Table 3 reduces the memory utilization.

3.2.4. Advantage of Fp_tree

Once the frequent pattern set created from transaction database D is huge, whereas its need to access the patterns regularly, it is possible to compress the sets into smaller one. To get frequent pattern easily, some patterns are reduce to enhance the performance. Given a set of frequent patterns $FP = \{fp_1 : s_1, fp_2 : s_2, \ldots fp_n : s_n\}$ where fp_i is a frequent pattern and s_i is support count. If there are two frequent patterns $fp_m : s_m$ and $fp_n : s_n$ where $s_m = s_n$ and $fp_m \in fp_n$ then pattern fp_m can be remove. The elimination can be applied only if the two frequent patterns are same and its support count are same. The elimination can also be applied if support count are same and the frequent item set is the subset of another frequent item set which has higher confidence. From Table 2 the frequent patterns {('bread':3), (jam:3), (bread jam butter:3), (bread butter:3)} is generated. The frequent pattern $(bread, butter)$ is removed from the data set, since the pattern 'bread butter' is the subset of (bread jam butter) with support count 3.

3.2.5. Construct FP-tree

Once the frequent items are converted and compressed the frequent pattern, the next step is to build FP-Tree. The entries in q^{th} row and r^{th} column of matrix M is M_{qr}. In the particular position the value is 1 which represents the q^{th} transactions in transaction database and r^{th} column in the Frequent_list. Matrix entries which has zero is not considered for constructing the tree. Once the matrix is constructed, the initial transaction database is not required for further processing. Hence it is erased from the memory which reduce the storage consumption level. The sample reduced constructed FP-Tree is depicted in Fig. 3. To simplify the tree traversal, an element in the header table is constructed. Items are subjected to its existence in the tree through a head of node link. Item in the same link is represented as node link in the tree. The conditional pattern base of the item is constructed with reference to these network paths. This process continues for entire data sets. Finally all frequent item sets are examined. Once the items are identified it is again decoded with corresponding entries with Frequent_list.

3.2.6. Apache's spark FP-Growth

Parallel enhanced FP-Growth helps to build FP-tree in a distributed machines using Apache's spark. Each node constructs the FP-tree individually and the final is merged to get global frequent itemsets. Based on the constructed conditional pattern tree the association rules are generated. FP-tree consists of a root node represented as null. A collection of item prefix sub tree is the children of the root, and a header table as represented as frequent-item. The prefix sub-tree contains three parts namely item-name, count and node-link. Item name represent the name of the item of the node. Count is represented as number of transaction in the path and node-link represent the next node link of a FP-tree. It has solid line from parent node to child node which represent the relationship between them. FP-Growth begins to mine the FP-tree on each item whose support count is greater than or equal to threshold to build the condition FP-tree as shown in Fig. 3. Though spark can process data in cache memory it improves the performance over MapReduce. It provides scalable and efficient data structure like RDD which describes an immutable group of elements operated parallel. Storing RDD in cache memory increase the performance of the system.These association rules helps to discover hidden pattern. Sample rules is given in Fig. 4. Large scale data problem is solved using Map-Reduce framework. But it requires lot of I/O for every operations. So it increase the computation cost to access data on every node.

3.3. Associative classification

Associative classification utilizes the association rule mining method which discover rules by examining highly useful rules that can simplify the training dataset. When compared to other classification technique the associative classification provides better results. To build the model for prediction association rule mining and classification modules are used.

3.4. Recommendation using association rule mining

Predicting user unknown preference are carried out in three process namely Rule discovery, Choosing strong rule and Classification.

3.4.1. Rule discovery

Based on association rules the rules are discovered form the training data set. These associative rules are called class associative rules.

3.4.2. Choosing strongest rule

The strongest rules were selected based on support and confidence of the rules which gives the accuracy of the classifier. The rules which do not satisfy the support threshold is eliminated and not consider for the classification.

3.4.3. Classification

The data is divided into training and test data. Association rules are generated from the training data. The rules are arranged in descending order based on confidence value. The strongest rules are selected from these training data for classification. In the prediction process, the test data is organized with highest confidence of the itemsets. These itemsets that are present in the test data where chosen for classification. Based on class label the prediction accuracy is calculated.

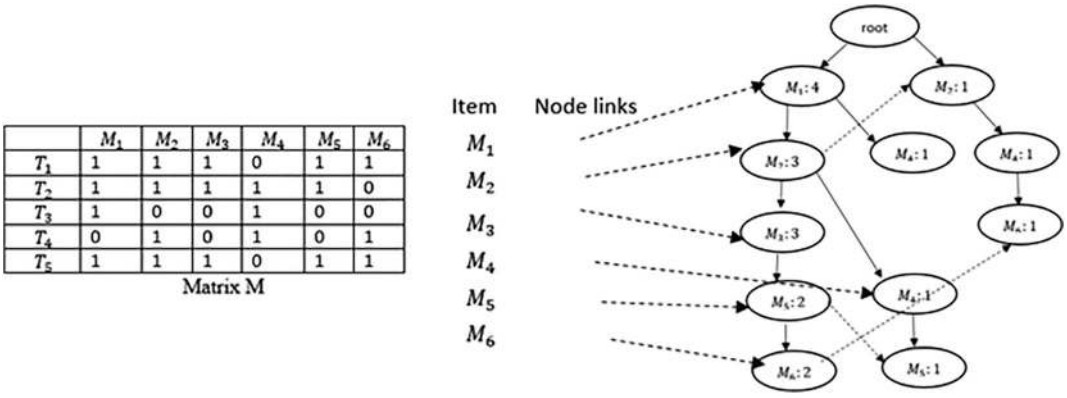
4. Experimental analysis

The famous benchmark dataset is used for experimental analysis from MovieLens 10 M[1] project.

[1] www.grouplens.org.

Table 3
Matrix Form.

	Frequent_list					
TiD	1	1	1	0	1	1
	1	1	1	1	1	1
	1	0	0	1	0	0
	0	1	0	1	0	1
	1	1	1	0	1	1



	M_1	M_2	M_3	M_4	M_5	M_6
T_1	1	1	1	0	1	1
T_2	1	1	1	1	1	0
T_3	1	0	0	1	0	0
T_4	0	1	0	1	0	1
T_5	1	1	1	0	1	1

Fig. 3. FP-Tree Construction.

No	Antecedent	Consequent	Support	Confidence	Lift
1	Thriller-Recent-High	Animation-Medium-Low, Children-Old-Middle, Fantasy-Medium-Middle, Comedy-Recent-High	0.7854	0.8506	1.6025
2	Adventures-Recent-High	Animation-Medium-Low, Fantasy-Recent-High, Comedy-Recent-High	0.7856	0.8516	1.5752
3	Children-Recent-High, Thriller-Medium-Low	Animation-Medium-Low, Fantasy-Recent-High, Comedy-Recent-High	0.7854	0.8528	1.5774
4	Action-Medium-Middle	Children-Old-Middle, Fantasy-Recent-High	0.7903	0.8550	0.9791
5	Animation-Medium-High, Thriller-Medium-High	Children-Middle-High, Fantasy-Medium-High, Comedy-Old-High	0.7954	0.8587	1.5746

Fig. 4. Generated rules.

4.1. Data set

The description of Movielens10M data set is furnished here. The study on data set shows that there are 15,220 unique tags. One tag is being used by a minimum of 4000 users. Each user rates approximately 142 movies with 10 unique tags. As a whole the dataset consists of 10 million ratings, 100,000 tag applied to 10,000 movies by 72,000 users.

4.2. Learning parameter

Support count and confidence are the important parameters to define the accuracy of the association rule mining approach. Based on the confidence, the strongest rules are selected from the generated list. When the support of the rule is very low its rule generation is high and vice versa. Choosing optimum value helps to find most frequent item set in the large data set. Various values were given to test the accuracy of the results and rule generation as illustrated in Figs. 5 and 6. Based on several trails the parameter was fixed as 0.8 and 0.7. Optimal support count eliminates the repeated rules and unnecessary rules. Optimal confidence value enhance the accuracy of the proposed method.

$$Precision = \frac{Number\ of\ recommended\ item\ that\ are\ relevant}{Number\ of\ recommended\ item} \quad (2)$$

$$Recall = \frac{Number\ of\ recommended\ item\ that\ are\ relevant}{Total\ number\ of\ relevant\ item} \quad (3)$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

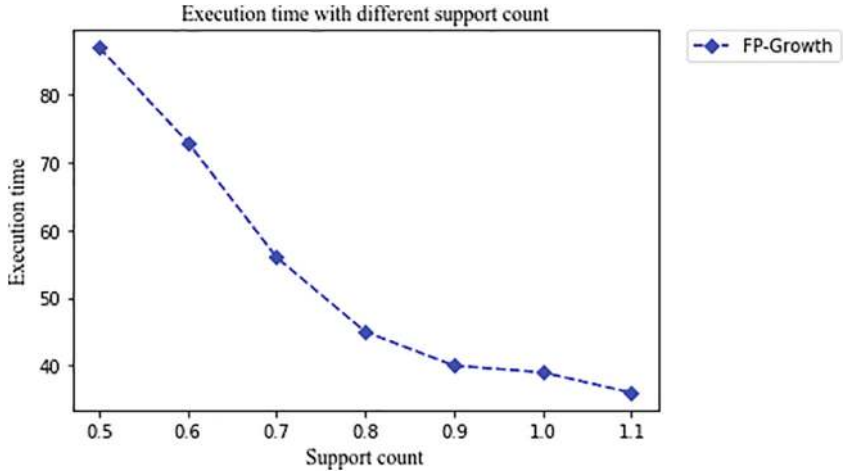


Fig. 5. Selecting support count.

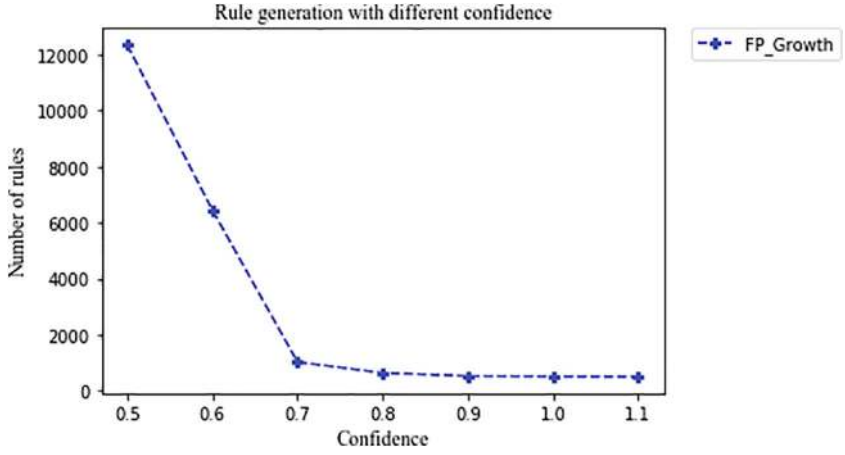


Fig. 6. Selecting confidence.

$$SparsityLevel = 1 - \frac{Number\ of\ ratings}{Number\ of\ users \times Number\ of\ items} \quad (5)$$

4.3. Evaluation metric

The main objective of the association rule mining is to extract the frequent item set from the transactional database. In some cases significant item set with low support are removed from this method. Support and Confidence metric were used in association rule mining. But to evaluate the performance of the proposed method, other metric such as precision, recall and f-score were taken into account. The data is split into training (80%) and testing (20%). The proposed method was trained with 80% training data set and generate the rules for training data set. The model was tested with testing data and generate the rules. Now it is necessary to compare the rules generated from testing and training. The precision and recall metric score were used to determine whether the items are related or unrelated (Lu et al., 2015). On the other hand it determines either suggested or not suggested (Yang et al., 2014). These scores are calculated using Eqs. (2) and (3). The precision score states how effective the method correctly selected the item for a recommendation, and the recall value states how many of those items are relevant. Both these metrics are inversely proportion to derive the optimal one (Bobadilla et al., 2013), F-score is suggested to solve the trade-off between these values. So f-score gives the average of precision and recall value as given in Eq. (4). The greater value of these metrics gives a greater accuracy level. The evaluation metric such as Support, Confidence and Lift are used to evaluate the association rule. But to compare with other traditional method the Precision, Recall and F-Score is calculated. The data sparsity level is calculated using the Eq. (5) and it is shown with the help of heat map in the Fig. 7 (Huang et al., 2016). The sparsity level ranges from 0–100% is taken into consideration for experimental purpose.

5. Result and discussion

The results obtained from the proposed method shows higher accuracy when compared to basic Collaborative Filtering (CF) method as shown in Fig. 8. An increase in the sparsity level decreases the performance of basic CF method. From the observations it is revealed that poor neighbourhood formation leads to poor recommendations. Hence the basic CF method suffers from sparsity problems. The proposed technique solves the data sparsity issue. Inaccurate neighbour selection in MovieLens data set brings unfruitful recommendations as prediction of preferences produces inaccurate results. From the Table 4 it is very clear that the proposed method has shown higher performance at various levels of sparsity. Moreover, while applying association rule mining to analyse the user movie watching pattern which depends on the genre

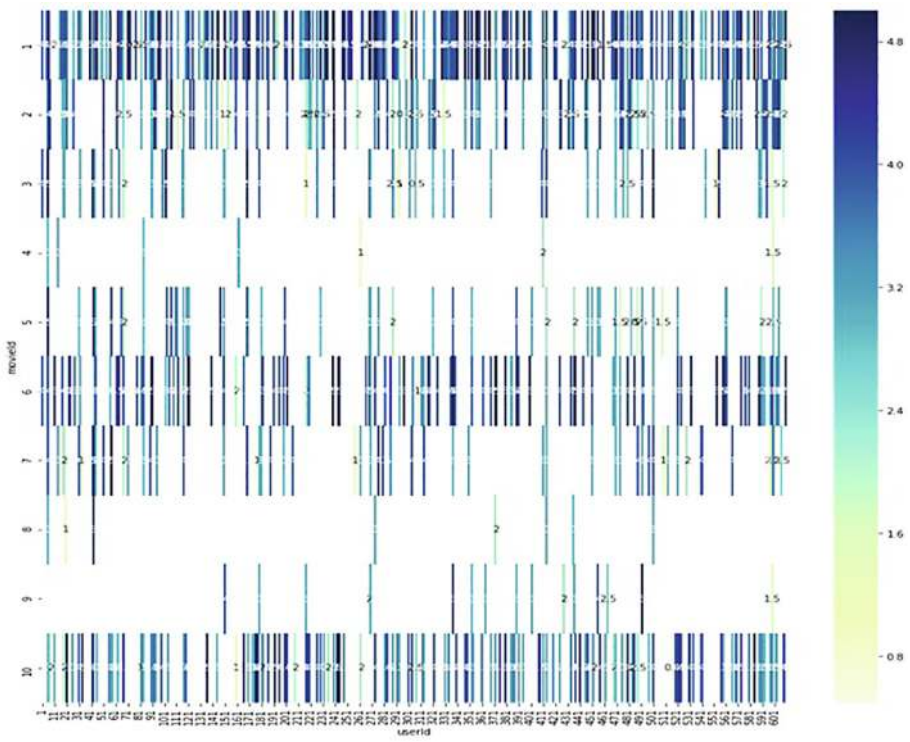


Fig. 7. Data Sparsity Level.

Data Sparsity	0-20(%)			20-40(%)			40-60(%)			60-80(%)			80-100(%)		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Basic CF	0.7012	0.6877	0.6943	0.6999	0.6755	0.6874	0.6756	0.6522	0.6312	0.6083	0.6066	0.6074	0.5978	0.5898	0.5937
Proposed Method	0.8077	0.7889	0.8032	0.8077	0.7989	0.8032	0.8077	0.7999	0.8032	0.7887	0.7702	0.7793	0.7787	0.7602	0.7793

Fig. 8. Comparison with basic CF method.

Table 4
Comparison of Existing Methods.

Different Models		Different Sparsity Levels(%)				
	Metrics	0–20	20–40	40–60	60–80	80–100
PCA + KMeans (Wang et al., 2014)	Precision	0.7525	0.7525	0.7422	0.7022	0.6556
	Recall	0.7622	0.7622	0.7534	0.7089	0.7001
	F-Score	0.7573	0.7573	0.7478	0.7055	0.6771
TriFac (Bao et al., 2012)	Precision	0.7214	0.7189	0.6978	0.6578	0.6078
	Recall	0.7123	0.7033	0.6933	0.6733	0.6833
	F-Score	0.7168	0.7110	0.6955	0.6655	0.6433
Proposed method	Precision	0.8177	0.8177	0.8077	0.7887	0.7787
	Recall	0.8089	0.8089	0.7989	0.7702	0.7602
	F-Score	0.8099	0.8099	0.8032	0.7793	0.7693

pattern, tag data that user has posted towards the item, and the recent user interest based on time. It not only predicts the preference score but it also analyse the user behavior. To analyse the popularity and pattern of movies that the user show their interest and to assess the hidden pattern from the user and an item rating matrix, the data is treated as equal. So the researcher consider using association rule mining as the best choice, when compare to other traditional methods. The association rule mining helps to discover set of items that occur together frequently in a user and item matrix. Its main objective is to identify group of items that are highly correlated with each other, or with respect to certain target variable. It works like a feature selection method.

5.1. Experiment 1 (Execution time)

Fig. 9(a) illustrates the comparison of the proposed FP-Growth with sequential FP-Growth and Apriori algorithm. The performance of the proposed method is tested with traditional Apriori algorithm with different support count which ranges from 0. 6% to 1%. It is noted that FP-Growth performs well when compared to other methods. When the support count is low, the methods yields more number of frequent items set and association rules. The Apriori algorithm generates huge number of candidate item sets. Hence, searching the pattern through huge candidates set becomes very expensive. The proposed methods required less time

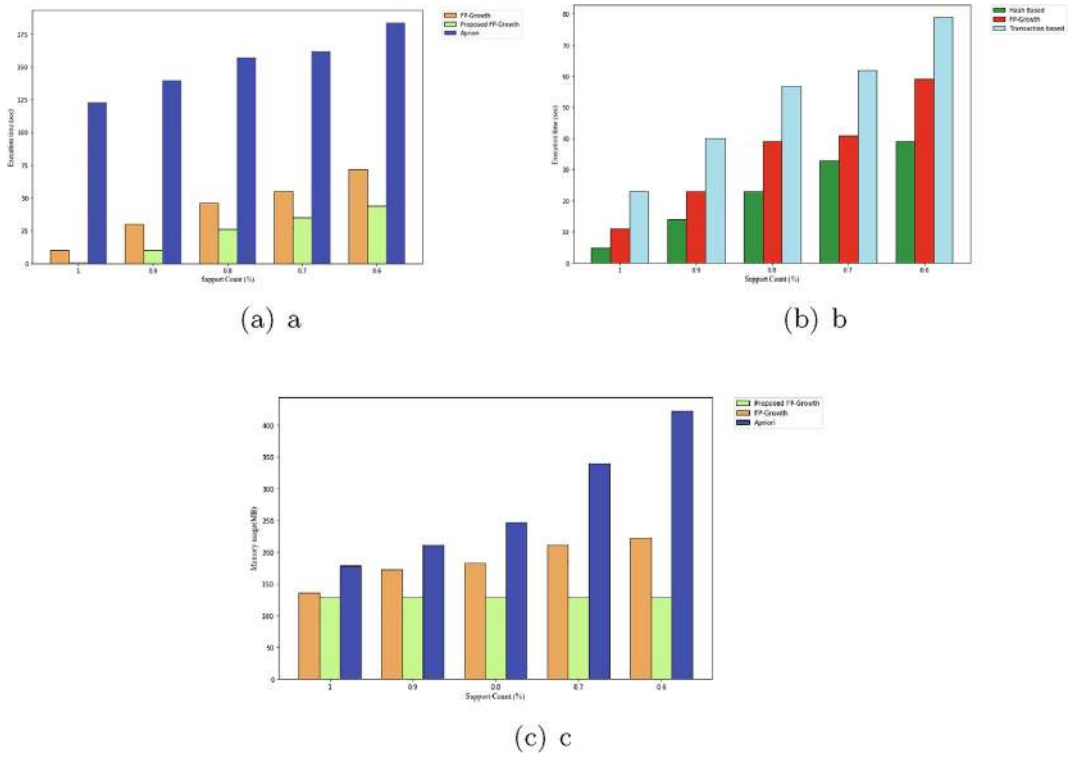


Fig. 9. (a) Execution time (b) Data structures (c) Memory utilization.

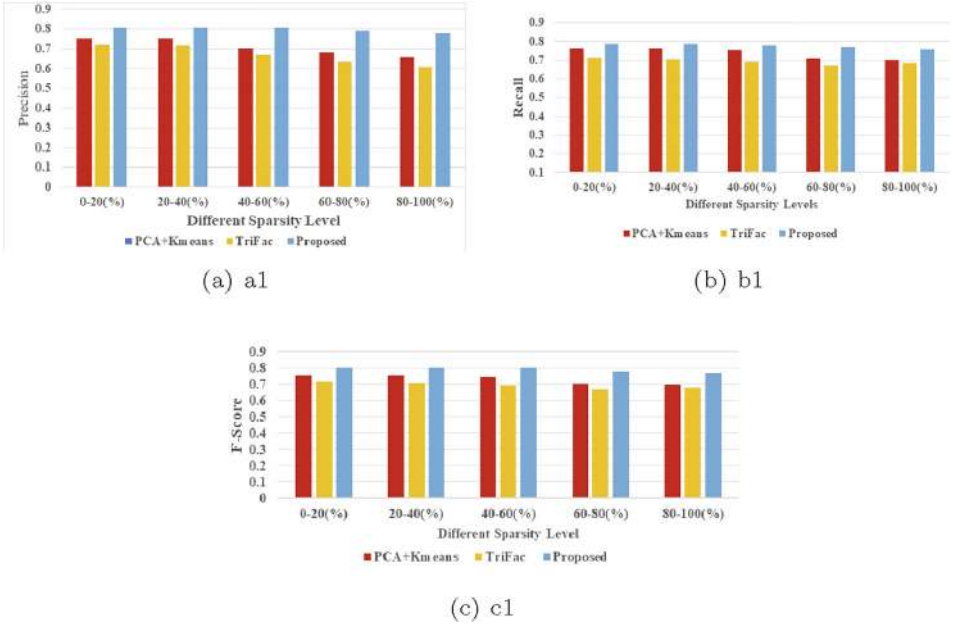


Fig. 10. (a1) Execution time (b1) Data structures (c1) Memory utilization.

when compared to other methods and achieves better performance.

5.2. Experiment 2 (Different data structure)

The efficiency of the association rule mining is tested with various data structures namely partition based, hash based and transaction based methods (Han et al., 2011). For experimental purpose, partition based and hash based methods were used for analysis. In Hash based method, hashing the itemset into corresponding buckets reduce the size of the candidate itemset generation (Park et al., 1995). Apriori with hashing is an additional technique which enhance the performance of Apriori algorithm. Partition based method permits parallel processing (Chung and Luo, 2008). It decrease the computation cost and enhance the performance of the association rule mining process. Apriori algorithm is an example of transaction based method. The tests were conducted based on various support count measures and the execution time for each

algorithm was observed. The aggregated results illustrated in Fig. 9 (b) discloses that the proposed FP-Growth achieves better results than Apriori with hashing and Apriori with partition based methods. Further the test was conducted to observe number of frequent items generated with respect to different support count. After first scan, the number of frequent item sets are observed with execution time and memory utilization. The basic Apriori based methods generate huge candidate sets. In hashing technique, repeated candidate item set generation is reduced to reasonable number. Hashing based method works well for small data sets, but become complex when it handles very large data sets. However the overall performance of proposed FP-Growth is high. The memory consumption for different data structures are illustrated in Fig. 9(c) reveals that proposed FP-Growth requires less time when compared to other methods.

5.3. Time complexity analysis

Let as assume that the transaction contains n items. It requires n × k times for arranging items. The searching time is k and comparing time is $k \times k$, then the complexity time is $O((n+k) \times k) = O(k^2)$. For apriori algorithm the time complexity is calculated as $O(2^i)$where i is referred as number of distinct items in transaction database.

6. Comparative analysis with existing method

Two methods were taken from the literature survey to evaluate the proposed methods.The results obtained from the experiment show that the presented method increases the recommendation accuracy in terms of precision, recall, and F-score is illustrated in Fig. 10(a), (b) and (c) respectively. The proposed method is compared with other CF based methods. Existing study on PCA + K means method solves the data sparsity issue using dimensionality reduction techniques such as PCA and predict the unknown preference using KMeans clustering methods (Wang et al., 2014). Using this method, it suffers with some information loss and it requires high computation cost when it handle large data. But the proposed method solves data sparsity problem by applying two-level clustering techniques. It predicts the user unknown preference by analysing the hidden pattern using association rules. The TriFac (Bao et al., 2012) model works under the principle of Probabilistic matrix factorization method and finds the latent features emphasis association among the user, item, rating, and tag. This method failed to handle overlapping condition where more than two tags were associated with an item. But the proposed method effective handles both overlapping condition and more than one tag associated with items. It increase the precision value 5% when compared to other methods. Recall values denote number of recommended items that are relevant. It shows that it enhances the recommendation accuracy. From the results observed, the experiment indicates that it works well at various data sparsity levels. Fig. 10(c) shows the results obtained from F-Score which is the average of precision and recall value listed in Table 4. The results show that the user preference is not only based on the rating score but also it is necessary to deeply analyze the hidden pattern knowledge. Hence the analyzing behavior of the user preferences enhances the recommendation accuracy.

7. Conclusion

Applying association rule helps to analyse the user interest, hidden pattern and correlations among preferred items.The method is tested with MovieLens Data Set which is a bench mark data set. Proposed pattern mining approach reduce the execution time by parallel processing and computation cost by storing frequent items in matrix form. Different sparsity levels were considered for experimental purpose and the results obtained from the experiments show that the proposed method outperforms various data sparsity level and generate recommendations even with high sparse data. Test results were compared with other CF based methods. The results obtained from experiment illustrate that the method achieves on average of 5% higher precision value, 3% higher recall value, and F-score on an average of 4% higher when compared to traditional CF method. The significant benefits of the presented method is to enrich the user profile by grouping items features such as items tag information, items category and the novelty of the object based on time. Further the method acquires hidden knowledge from the item preference to predict the unknown preferences. The advantage of proposed method analyse most frequently occurring items that reveals hidden correlation, association and pattern behind the preferred items. Instead of recommending popular items, the proposed method analyse users hidden interest and recommend item based on the users previous interest pattern. As the interest of users are dynamic, the proposed method analyse and rank the interest based on recent preference and will not recommend the outdated items. Results observed from the experiment shows that 83% of users likes “Drama” Movies with two or more genre like Animation, Adventure and Children.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Bao, T., Ge, Y., Chen, E., Xiong, H., Tian, J., 2012. Collaborative filtering with user ratings and tags. In: Proceedings of the 1st International Workshop on Context Discovery and Data Mining. ACM, p. 1.

Belhadi, A., Djenouri, Y., Lin, J.C.-W., Cano, A., 2020a. A data-driven approach for twitter hashtag recommendation. IEEE Access 8, 79182–79191.

Belhadi, A., Djenouri, Y., Lin, J.C.-W., Cano, A., 2020b. A general-purpose distributed pattern mining system. Applied Intelligence, 1–16.

Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A., 2013. Recommender systems survey. Knowledge-Based Systems 46, 109–132.

Chung, S.M., Luo, C., 2008. Efficient mining of maximal frequent itemsets from databases on a cluster of workstations. Knowledge and Information Systems 16 (3), 359–391.

Djenouri, Y., Lin, J.C.-W., Nørvåg, K., Ramampiaro, H., 2019. Highly efficient pattern mining based on transaction decomposition. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE, pp. 1646–1649.

Han, J., Pei, J., Kamber, M., 2011. Data Mining: Concepts and Techniques. Elsevier.

Huang, T.C.-K., Chen, Y.-L., Chen, M.-C., 2016. A novel recommendation model with google similarity. Decision Support Systems 89, 17–27.

Isinkaye, F., Folajimi, Y., Ojokoh, B., 2015. Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal 16 (3), 261–273.

Kumar, S., Mohbey, K.K., A review on big data based parallel and distributed approaches of pattern mining. Journal of King Saud University-Computer and Information Sciences.

Kumar, P., Thakur, R.S., 2018. Recommendation system techniques and related issues: a survey. International Journal of Information Technology 10 (4), 495–501.

Lin, X., 2014. Mr-apriori: Association rules algorithm based on mapreduce. In: 2014 IEEE 5th International Conference on Software Engineering and Service Science, IEEE, 2014, pp. 141–144.

Lu, J., Wu, D., Mao, M., Wang, W., Zhang, G., 2015. Recommender system application developments: a survey. Decision Support Systems 74, 12–32.

Luna, J.M., Fournier-Viger, P., Ventura, S., 2019. Frequent itemset mining: A 25 years review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (6). e1329.

Manogaran, G., Varatharajan, R., Priyan, M., 2018. Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. Multimedia Tools and Applications 77 (4), 4379–4399.

Maulik, U., Bandyopadhyay, S., 2002. Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (12), 1650–1654.

Miao,Y., Lin, J., Xu, N., 2019. An improved parallel fp-growth algorithm based on spark and its application. In: 2019 Chinese Control Conference (CCC), IEEE, pp. 3793–3797.

Moradi, P., Rezaimehr, F., Ahmadian, S., Jalili, M., 2016. A trust-aware recommender algorithm based on users overlapping community structure. In: 2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE, pp. 162–167.

Najafabadi, M.K., Mahrin, M.N., Chuprat, S., Sarkan, H.M., 2017. Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data. Computers in Human Behavior 67, 113–128.

Padillo, F., Luna, J.M., Ventura, S., 2020. Lac: Library for associative classification. Knowledge-Based Systems 193. 105432.

Park, J.S., Chen, M.-S., Yu, P.S., 1995. An effective hash-based algorithm for mining association rules. Acm Sigmod Record 24 (2), 175–186.

Sethi, K.K., Ramesh, D., 2017. Hfim: a spark-based hybrid frequent itemset mining algorithm for big data processing. The Journal of Supercomputing 73 (8), 3652–3668.

Solanki, S.K., Patel, J.T., 2015. A survey on association rule mining. In: 2015 Fifth International Conference on Advanced Computing & Communication Technologies. IEEE, pp. 212–216.

Tyagi, S., Bharadwaj, K.K., 2013. Enhancing collaborative filtering recommendations by utilizing multi-objective particle swarm optimization embedded association rule mining. Swarm and Evolutionary Computation 13, 1–12.

Wang, Z., Yu, X., Feng, N., Wang, Z., 2014. An improved collaborative movie recommendation system using computational intelligence. Journal of Visual Languages & Computing 25 (6), 667–675.

Wu, B., Zhang, D., Lan, Q., Zheng, J., 2008. An efficient frequent patterns mining algorithm based on apriori algorithm and the compact -tree structure. In: 2008 Third International Conference on Convergence and Hybrid Information Technology, vol. 1, IEEE, pp. 1099–1102.

Yang, X., Guo, Y., Liu, Y., Steck, H., 2014. A survey of collaborative filtering based social recommender systems. Computer Communications 41, 1–10.

Yang, S., Xu, G., Wang, Z., Zhou, F., 2015. The parallel improved apriori algorithm research based on spark. In: 2015 Ninth International Conference on Frontier of Computer Science and Technology. IEEE, pp. 354–359.