A NEW HYBRID ALGORITHM FOR BUSINESS INTELLIGENCE RECOMMENDER SYSTEM

P.Prabhu[1] and N.Anbazhagan[2]

[1]Directorate of Distance Education, Alagappa University, Karaikudi, Tamilnadu, INDIA

[2]Department of Mathematics, Alagappa University, Karaikudi, Tamilnadu, INDIA

ABSTRACT

Business Intelligence is a set of methods, process and technologies that transform raw data into meaningful and useful information. Recommender system is one of business intelligence system that is used to obtain knowledge to the active user for better decision making. Recommender systems apply data mining techniques to the problem of making personalized recommendations for information. Due to the growth in the number of information and the users in recent years offers challenges in recommender systems. Collaborative, content, demographic and knowledge-based are four different types of recommendations systems. In this paper, a new hybrid algorithm is proposed for recommender system which combines knowledge based, profile of the users and most frequent item mining technique to obtain intelligence.

KEYWORDS

Business Intelligence, Frequent Itemset , k-means Clustering, Data Mining, Decision Making, Recommender system, E-commerce.

1. INTRODUCTION

Data Base Management System (DBMS) and Data Mining (DM) are two emerging technologies in this information world. Knowledge is obtained through the collection of information. Information is enriched in today's business world. In order to maintain the information, a new systematic way has been used such as database. In this database, there are collection of data organized in the form of tuples and attributes. In order to obtain knowledge from a collection of data, business intelligence methods are used. Data Mining is the powerful new technology with great potential that help the business environments to focus on only the essential information in their data warehouse. Using the data mining technology, it is easy for decision making by improving the business intelligence.

Frequent itemset mining is to find all the frequent itemsets that satisfy the minimum support and confidence threshold. Support and Confidence are two measures used to find the interesting frequent itemsets. In this paper, frequent itemset mining can be used to search for frequent item set in the data warehouse. Based on the result, these frequent itemsets are grouped into clusters to identify the similarity of objects.

Cluster Analysis is an effective method of analyzing and finding useful information in terms of grouping of objects from large amount of data. To group the data into clusters, many algorithms have been proposed such as k-means algorithm, Fuzzy C means, Evolutionary Algorithm and EM Method. These clustering algorithms groups the data into classes or clusters so that object within

a cluster exhibit same similarity and dissimilar to other clusters. Thus based on the similar-ity and dissimilarity, the objects are grouped into clusters.
In this paper, a new hybrid algorithm is for business intelligence recommender system based on knowledge of users and frequent items. This algorithms works in three phases namely pre-processing, modelling and obtaining intelligence. First, the users are filtered based on the user's profile and knowledge such as needs and preferences defined in the form of rules. This poses selection of features and data reduction from dataset. Second, these filtered users are then clustered using k-means clustering algorithm as a modelling phase. Third, identifies nearest neighbour for active users and generates recommendations by finding most frequent items from identified cluster of users. This algorithm is experimentally tested with e-commerce application for better decision making by recommending top n products to the active users.

2. RELATED WORKS

Alexandre et al, [1] presented a framework for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve privacy of individual transactions. They analyzed the nature of privacy breaches and proposed a class of randomization operators that are much more effective than uniform randomization in limiting the breaches.

Jiaqi Wang et al, [4] stated that Support vector machines (SVM) have been applied to build classifiers, which can help users make well-informed business decisions. The paper speeds up the response of SVM classifiers by reducing the number of support vectors. It was done by the K-means SVM (KMSVM) algorithm proposed in the paper. The KMSVM algorithm combines the K-means clustering technique with SVM and requires one more input parameter to be determined: the number of clusters..

M. H. Marghny et al, [5], stated that Clustering analysis plays an important role in scientific research and commercial application. In the article, they proposed a technique to handle large scale data, which can select initial clustering center purposefully using Genetic algorithms (GAs), reduce the sensitivity to isolated point, avoid dissevering big cluster, and overcome deflexion of data in some degree that caused by the disproportion in data partitioning owing to adoption of multi-sampling.

Wenbin Fang et al, [11], presented two efficient Apriori implementations of Frequent Itemset Mining (FIM) that utilize new-generation graphics processing units (GPUs). The implementations take advantage of the GPU's massively multi-threaded SIMD (Single Instruction, Multiple Data) architecture. Both implementations employ a bitmap data structure to exploit the GPU's SIMD parallelism and to accelerate the frequency counting operation.

Ravindra Jain, [8], explained that data clustering was a process of arranging similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group was better than among groups. In the paper a hybrid clustering algorithm based on K-mean and K-harmonic mean (KHM) was described. The result obtained from proposed hybrid algorithm was much better than the traditional K-mean & KHM algorithm.

David et al, [3], described a clustering method for unsupervised classification of objects in large data sets. The new methodology combines the mixture likelihood approach with a sampling and sub sampling strategy in order to cluster large data sets efficiently. The method was quick and reliable and produces classifications comparable to previous work on these data using supervised clustering.

Risto Vaarandi, [9], stated that event logs contained vast amounts of data that can easily overwhelm a human. Therefore, mining patterns from event logs was an important system management task. The paper presented a novel clustering algorithm for log file data sets which helps one to detect frequent patterns from log files, to build log file profiles, and to identify anomalous log file lines.

R. Venu Babu and K. Srinivas [10] presented the literature survey on cluster based collaborative filter and an approach to construct it. In modern E-Commerce it is not easy for active users to find the best suitable goods of their interest as more and more information is placed on line (like movies, audios, books, documents etc...). So in order to provide most suitable information of high value to active users of an e-commerce business system, a customized recommender system is required. Collaborative Filtering has become a popular technique for reducing this information overload. While traditional collaborative filtering systems have been a substantial success, there are several problems that researchers and commercial applications have identified: the early rater problem, the sparsity problem, and the scalability problem.

3. NEW HYBRID ALGORITHM

In this business world, there exists a lot of information. It is necessary to maintain the information for decision making in business environment. The decision making consists of two kinds of data such as OnLine Analytical Processing (OLAP) and OnLine Transactional Processing (OLTP).The former contains historical data about the business from the beginning itself and the later contains only day-to-day transactions on business. Based on these two kinds of data, decision making process can be carried out by means of a new hybrid algorithm based on frequent itemsets mining and clustering using k-means algorithm and knowledge of users in order to improve the business intelligence. The proposed new hybrid algorithm design is shown in figure 1.

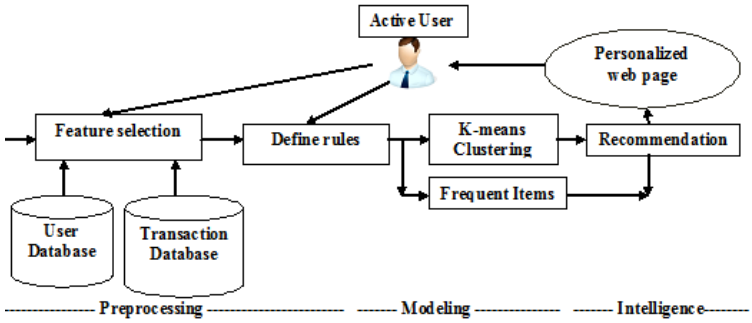

Figure 1. A New Hybrid algorithm design

The step involved in the proposed Hybrid algorithm is given below:

- Identifying the dataset
- Choose the consideration columns/features
- Filtering objects by defining the rules
- Identifying frequent items
- Cluster objects using k-means clustering
- Find nearest neighbour of active user
- Generate recommendation dataset for active user.

3.1. Identifying the dataset

To maintain the data systematically and efficiently, database and data warehouse technologies are used. The data warehouse not only deals with the business activities but also contains the information about the user that deals with the business. The representation of the data set is shown below:

$$D=\Sigma (A) = \{a_1, a_2,\ldots\ldots\ldots\ldots, a_n\} \quad (1)$$

Where (A) is the collection of all attributes, a_1, a_2, are the attribute list that deals with the dataset. Upon collecting the data, the dataset contains the data as follows:

$$\begin{pmatrix} a_{00} & a_{01} & a_{02}\ldots\ldots\ldots\ldots a_{0n} \\ a_{10} & a_{11} & a_{12}\ldots\ldots\ldots\ldots a_{1n} \\ . & & \\ . & & \\ . & & \end{pmatrix}$$

Here, a_{ij} is the data elements in the dataset, where i= 0,1,…n and j= 0,1,…m.

3.2. Choosing the considering columns/features

Upon the dataset has been identified, the next step of the proposed work is to choose the consideration column or filtering columns/features. That is, from the whole dataset, the columns/subset of features to be considered for our work has been chosen. This includes the elimination of the irrelevant column in the dataset. The irrelevant column/feature may the one which provide less information about the dataset.

$$CC = \Sigma (A`) = \Sigma (A) - \Sigma (A_u) \quad (2)$$

$$\Sigma (A`) = \{a_1, a_2, \ldots, a_n\} - \{a_{u1}, a_{u2},\ldots,a_{un}\} \quad (3)$$

$$CC = D`=\sum(A') = \{a'_1, a'_2, a'_3, \ldots, a'_m\} \quad (4)$$

Where, CC denotes the consideration column, which will be represented as Σ (A`),Σ (A) represents the set of all attributes in the chosen dataset, Σ (A_u) represents the set of all features to be eliminated to get the subset of features. The consideration column of the dataset can be represented as follows:

$$\begin{pmatrix} a`_{00} & a`_{01} & a`_{02}\ldots\ldots\ldots a`_{0n} \\ a`_{10} & a`_{11} & a`_{12}\ldots\ldots\ldots a`_{1n} \\ & & . \\ & & . \\ & & . \\ -` & -` & -` \end{pmatrix}$$

Here, $a`_{ij}$ is the data elements in the new resultant dataset with consideration column, where i = 0, 1,…, n and j = 0,1,…m.

3.3. Filtering objects by defining rules

From the consideration dataset, the objects can be grouped under stated conditions that are defined in terms of rules. That is, for each column that is considered, specify the rule to extract the necessary domain from the original dataset. This rule is considered to be the threshold value. The domain can be chosen by identifying the frequent items from the dataset. Rule can be defined as;

$$R = \{x/x \in D', x \geq 20 \text{ and } x \leq 60\} \quad (5)$$

$$\sum(R) = \sum(R) \bowtie \sigma((a_{ij}(A')/D') \geq 20 \text{ and } (a_{ij}(A')/D') \leq 60)$$

Where,

R	denotes Rule
σ	denotes the selection
$\bowtie$	denotes join
$a_{ij}(A')$	denotes the attribute from the dataset D′
D′	denotes dataset with selected attributes.

3.4 Identifying frequent items

The frequent items can be identified by analyzing the repeated value in the consideration column

$$FIS = value(S) > (SUP(S) \text{ and/or } CONF(S)) \quad (6)$$

Where, FIS represents the identified frequent itemset. Value(S) is the frequent items in the column S, satisfying the SUP(S) and CONF(S).SUP(S) is defined as the percentage of objects in the dataset which contain the item set. CONF(S) is defined as SUP(X U Y) / SUP (X). (i.e.,) the confidence on the frequent item set can be determined by combining the X and Y values from the dataset and then neglecting the X value to obtain the frequent item. Any objects that satisfy the criteria are selected and counted. This can be carried out by:

$$C_n = \square (a_{ij}(A) / D`) \quad (7)$$

It counts the number of domains, a_{ij} of the attribute list, A from the Dataset, D. From the counted value, C_n, we can determine the frequent item set that has been occurred in the dataset using the threshold value T.

$$C_n (a_{ij}(A)) > T \quad (8)$$

It shows that the domain a_{ij} of attribute A satisfies the threshold value T specified.

3.5 Clustering objects/users using k-means clustering

Upon forming the new dataset D``, the objects in D`` are clustered based on similarity of objects using k-means clustering. k-means clustering is a method of classifying or grouping objects into k clusters (where k is the number of clusters). The clustering is performed by minimizing the sum

of squared distances between the objects and the corresponding centroid. The resultant consists of cluster of objects with their labels/classes.

3.6 Find the nearest neighbour of active user

In order to find the nearest neighbours of the active user, similarity of the active user between cluster centroids are calculated based on distance measure. Then, select cluster that have the highest similarity among other clusters.

3.7 Generate recommendation dataset for active user

Recommendations are generated for the active user based on the selected cluster of users purchased most frequent items generated from specified threshold T. This gives intelligence to the users and business for better decision making.

3.8 New Hybrid Algorithm

Algorithm: New Hybrid Algorithm
Input:
The number of clusters *k* .
Dataset D with *n* objects.

Output: A set of clusters C_k.
Begin
Identify the dataset $D = \Sigma (A) = \{a_1, a_2, \ldots\ldots\ldots, a_n\}$attributes/objects.
Outline the Consideration Column (CC) from D.
$CC = D` = \sum(A') = \{a'_1, a'_2, a'_3, \ldots, a'_m\}$
repeat
Formulate the rules for identifying the similar objects.
$\sum(R) = \sum(R) \bowtie \sigma(a_{ij}(A')/D')$, where $i = 1$ to n , $j = 1$ to m .
S = f(X) / D, where S is the sample set containing identified column
FIS = value(S) > (SUP(X) and/or (SUP(X U Y) / SUP (X))),
where FIS is the frequent itemsets identified.
C_n (a_{ij} (A`)) > T from FIS, where T specifies the threshold value.
Generate the Resultant Dataset, D``
until no further partition is possible in CC.
Identify the *k* initial mean vectors (centroids) from the objects of D``.
Repeat
Compute the distance, □ between the object a_i and the centroids c_j.
Assign objects to cluster with min{∂(c_{jk})} of all clusters
Recalculate the *k* new centroids c_j from the new cluster formed
Until reaching convergence
Find the nearest neighbour of active object
Generate recommendation from most frequent items of nearest neighbour
End

4. EXPERIMENTAL SETUP

The proposed hybrid algorithm provides solution for recommender systems. This methodology can be verified through various experimental setups. In this work E-Commerce dataset is used for testing the proposed algorithm for recommending products purchase by the users. This algorithm

help active users find items they want to buy from a business. The E-commerce business, using recommender systems are Amazon.com, CDNOW.com, Drugstore.com, eBay, MovieFinder.com and Reel.com. The Table 1 shows the description of sample dataset.

Table 1 Dataset Description

Key Element	**Description**
Dataset Name	E-Commerce Synthetic dataset 1
Original Attribute-list	Age, Gender, Occupation, Salary, Date, Time, item ,rate, amount, rating
Consideration Column	Age, Gender, Occupation, Salary, Amount
Rules defined	Age >=21 and <= 25 Occupation = “Teacher” Amount > 25000

5 RESULTS AND DISCUSSION

Based on the identified frequent item set mining, it is clear that the user with age between 21 and 25, frequently accessing the site. And so, the frequent item set being the user with age less than 25 and greater than 21. From the original dataset, proposed method identified the products purchased by the users belonging to these age group is shown in table 2.

Table 2 Frequent Items Identified

Age/ user-id	**Item id**	**Frequently Items**
21	1010	Computer
22	2001	Jewels
23	3001	Books
24	5010	Sports
25	4101	Shoes

The identified frequent items based on the defined rules are counted to form the resultant dataset D``. Now k-means clustering is applied to group the users based on similarities. The table 3 shows the initial centroids.

Table 3 Initial Centroids

Cluster	**Age**	**Mean Vector (Centroid)**
1	21	(1,1)
2	23	(5,7)

The resulting objects found in the cluster -1 are 21 and 21. The objects found in the cluster-2 rare 23, 24 and 25. The table 4 shows the distance between the object to their centroid of each clusters.

Table 4 The distance between the objects to the centroid of each cluster.

Age/user id	Distance between the object to the centroid in Cluster-1	Distance between the object to the centroid in Cluster-2
21	1.8	5
22	1.8	2
23	5.4	3
24	4.0	2.2
25	2.1	1.4

Each individual's distance to its own cluster mean should be smaller that the distance to the other cluster's mean. Thus the mean of the object 21, 22 belongs to cluster-1 which is nearer to cluster-1, whereas the mean of the object 23, 24, 25 belongs to cluster-2, which is nearer to cluster-2. Thus there is no relocation occur in this example. From the obtained results, it is clear that the users are grouped under 2 clusters named as cluster-1 and cluster-2.

When the active user enters into the site, the first step is to verify the active user details to identify, on which cluster the user can be fall on. This can be done by analyzing the nearest neighbour of active user with existing clusters. Based on the distance, the active user can be easily classified and identified their position on the clusters. In our experiment, if the active user belongs to age 22, then they fall under cluster-1 and then we conclude that this user can have more probability to purchase either computer or jewels. Thus the upcoming user can now be recommended and redirected to the web page containing the details of computers and jewels. Through this kind of redirection, it is easy for the active user to save the time to search for their desired product. Thus, by grouping the similar behavior users into a cluster and based on the cluster result, the active user can be recommended and redirected to that products web page.Hence this intelligence provides active users / business for better decision making by recommending the top products.

The figure 2 shows sample cluster of users generated using proposed clustering approach for k =5 with 41 users of synthetic dataset 2.

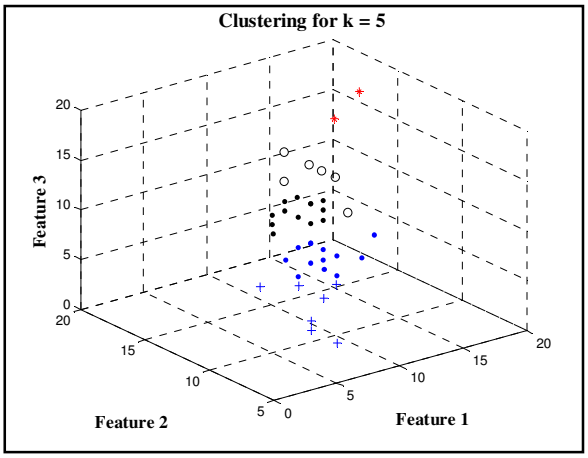

Figure 2. Sample cluster of users for k =5

The table 5 shows the allocation of users in the clusters for k =5 using proposed method.

Table 5 Allocated user Id's in the clusters for k =5 of synthetic dataset-2.

Cluster Id	**Allocated user id's**	**Total**	**Centroid**	**Centroid values**		
1	21,23,39,40,46, 52,55	7	C_1	17.42	16.45	16.41
2	20,22,24,27,30, 31,34,44,45,54, 56.57,59	13	C_2	4.88	6.67	6.16
3	41,50	2	C_3	11.90	13.94	12.80
4	28,29,32,42,48, 49	6	C_4	9.59	11.87	11.40
5	25,26,33,35,36, 37,38,43,47,51, 53,58,60	13	C_5	8.40	9.82	8.51

Table 6 shows the sample of identified frequent items purchased by the users.

Table 6 Sample of Identified frequent items

	Support		
	50%	**60%**	**70%**
User Id	**Item id**	**Item id**	**Item id**
20	101,310,401	101,310	101
21	102,308	308	308
22	105	105	105
23	219,207	219	219
24	101,309	101,309	101
-	-	-	-
-	-	-	-
41	101,209	101,209	209
-	-	-	-
40	103,310,311	103,311	311
60	123	123	123

The experimental results of frequent items identified using various support count is presented. The active users, identified clusters and recommended items using k=10 and support =50% are tabulated in Table 7.

Table 7 Recommended items of active user

Active user	**Identified Cluster Id**	**Allocated User id's**	**Recommended Item Id's**		
			Support 50%	**Support 60%**	**Support 70%**
AU_1	3	41,50	101,209,103, 310,311	101,209, 103,311	209,311

In this example, Cluster Id 3 is identified as neighbour for the active user 1 AU_1. Hence frequent items (101,209,103,310,311) from the objects in the cluster Id 2 are recommended for the active

user. This gives intelligence to the active users for selecting their items based on their profile and preferences and improves business activities.

Also, we can compare the performance of our proposed algorithm with the existing methodologies like [10] with various metrics like precision,recall and silhoutte index. This can be carried out by finding quality with various number of neighbours,iterations and clusters. The performance of the proposed method performs better than existing methods.

6. CONCLUSIONS AND FUTURE WORK

Business intelligence is a new technology for extracting information for a business from its user databases. In this paper we presented and evaluated new hybrid algorithm for improving business intelligence for better decision making by recommending products purchase by the user. The performance of the methodology can be verified by undertaking many experimental setups. The results obtained from the experiments shows that the methodology performs well. This algorithm can be tested with many real-world datasets with different metrics as a future work.

REFERENCES

[1] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, Johannes Gehrke, "Privacy Preserving Mining of Association Rules", IBM Almaden Research Center, USA, Copyright 2002.

[2] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl,Item-Based Collaborative Filtering Recommendation Algorithms,Procedings of the 10 th WWW International Conference, , Hong Kong,ACM 1-58113-348-0/01/0005.p285-295 May 1-5, 2001.

[3] David M. Rocke and Jian Dai, "Sampling and Subsampling for Cluster Analysis in Data Mining: With Applications to Sky Survey Data", Data Mining and Knowledge Discovery, 7, 215-232, 2003

[4] Jiaqi Wang, Xindong Wu, Chengqi Zhang, "Support vector machines based on K-means clustering for real-time business intelligence systems", Int. J. Business Intelligence and Data Mining, Vol. 1, No. 1, 2005.

[5] M. H. Marghny, Rasha M. Abd El-Aziz, Ahmed I. Taloba, "An Effective Evolutionary Clustering Algorithm: Hepatitis C Case Study", Computer Science Department, Egypt, International Journal of Computer Applications (0975-8887), Volume 34-No.6, Nov. 2011.

[6] P.Prabhu,' Method for Determining Optimum Number of Clusters for Clustering Gene Expression Cancer Dataset', International Journal of Advance Research in Computer Science pg 315(Volume 2 No. 4, July-August 2011).

[7] P.Prabhu, N.Anbazhagan, 'Improving the performance of k-means clustering for high dimensional dataset', International Journal of Computer Science and Engineering, Vol 3. No.6. Pg 2317-2322, June 2011.

[8] Ravindra Jain, "A Hybrid Clustering Algorithm for Data Mining", CCSEA, SEA, CLOUD, DKMP, CS & IT 05, pp. 387-393, 2012.

[9] Risto Vaarandi, "A Data Clustering Algorithm for Mining Patterns from Event Logs", Proceedings IEEE Workshop on IP Operations and Management, 2003.

[10] R. Venu Babu, K. Srinivas: A New Approach for Cluster Based Collaborative Filters. International Journal of Engineering Science and Technology Vol. 2(11), , 6585-6592, 2010.

[11] Wenbin Fang, Mian Lu, Xiangye Xiao, Bingsheng He, Qiong Luo, "Frequent Itemset Mining on Graphics Processors", Proceedings of the Fifth International Workshop on Data Management on New Hardware,. June 2009

[12] Zan Huang, Daniel Zeng and Hsinchun Chen,A Comparision of Collaborative-Filtering Recommendation Algorithms for E-commerce, IEEE INTELLIGENT SYSTEMS, IEEE Computer Society,p68-78,2007.