Using Affiliation Rules-based Data Mining Technique in Referral System

*Odule, Tola J. Adesina, Ademola O. Abdullah, K-K. Adebisi, **Ogunyinka, Peter I.**

Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye. Nigeria

Abstract

Referral techniques are normally employed in internet business applications. Existing frameworks prescribe things to a particular client according to client inclinations and former high evaluations. Quite a number of methods, such as cooperative filtering and content-based methodologies, dominate the architectural design of referral frameworks. Many referral schemes are domain-specific and cannot be deployed in a general-purpose setting. This study proposes a two-dimensional (User × Item)-space multimode referral scheme, having an enormous client base but few articles on offer. Additionally, the design of the referral scheme is anchored on the $Favourite$ and $Non-Favorite$ articles, as expressed by a particular client, and is a combination of affiliation rules mining and the content-based method. The experiments used the dataset of MovieLens, consisting of 100,000 motion pictures appraisals on a size of 1-5, from 943 clients on 1,682 motion pictures. It utilised a five-overlap cross appraisal on a (User × Item)-rating matrix with 12 articles evaluated by a minimum of 320 clients. A total of 16 rules were generated for both $Favourite$ and $Non-Favorite$ articles, at 35% minimum support and 80% confidence for the $Favourite$ articles and 50% similitude for the $Non-Favorite\ Items$. Experimental results showed that the anticipated appraisals in denary give a better rating than other measures of exactness. In conclusion, the proposed algorithm works well and fits on two dimensional $(User \times Item)$-space with articles that are significantly fewer than users, thus making it applicable and effective in a variety of uses and scenarios as a general-purpose utility.

Keywords: Referral system, Data mining techniques, Association rule mining, Apriori algorithm, Multimode referral system.

1.Introduction

A referral system can provide suggestions (recommendations) to users in multiple contexts, such as when they are choosing among an extensive collection of items. Referral systems strive to predict unrated items for a particular user [1]. More formally, let U be a lot of every single imaginable client, and let I be a lot of every single imaginable thing. Give f a chance to be a utility capacity that estimates the value of thing I to a client u; for example, $U \times I \to R$, where R is an arranged arrangement of non-negative whole numbers or genuine numbers. At that point, for every client $u \in U$, it is required to pick a thing $I_u \in I$ to boost the user's utility, as demonstrated as follows

$$\forall\, u\, \in U, i_u = \arg\max f(u, i) \tag{1}$$

With regards to referral frameworks, the utility of a thing is generally characterised by a rating. Based on that predicated ratings, the frameworks select things with the most elevated anticipated appraisals and prescribe them to the client.

Referral systems (RS) generally have four key features: prediction; individualised ranking; providing user feedbacks and; suggestion based on similarity. Referral engines collect different types of data;

however, whatever the data source is, three entities are generally identified: items, users and relations between users and items.

Experience goods identify assets that are consumed before knowing their satisfaction level. Shoppers confronted the troublesome errand of utilizing their constrained spending plans to obtain a portion of these substances, without completely realizing how satisfying they are. In such circumstances, referrals can offer a generous improvement in basic leadership of what to buy. The main objective of this study is to join affiliation rules mining and substance-based way to give a structure for a multimode referral system on a two-dimensional $(user \times item)$ space, with the proviso that the $(user \times item)$ space has a huge client base (> 1000) with relatively few offerings (< 50).

Data Mining Techniques for Referral Systems

The Figure below provides an overview of the Data Mining techniques used in this paper.
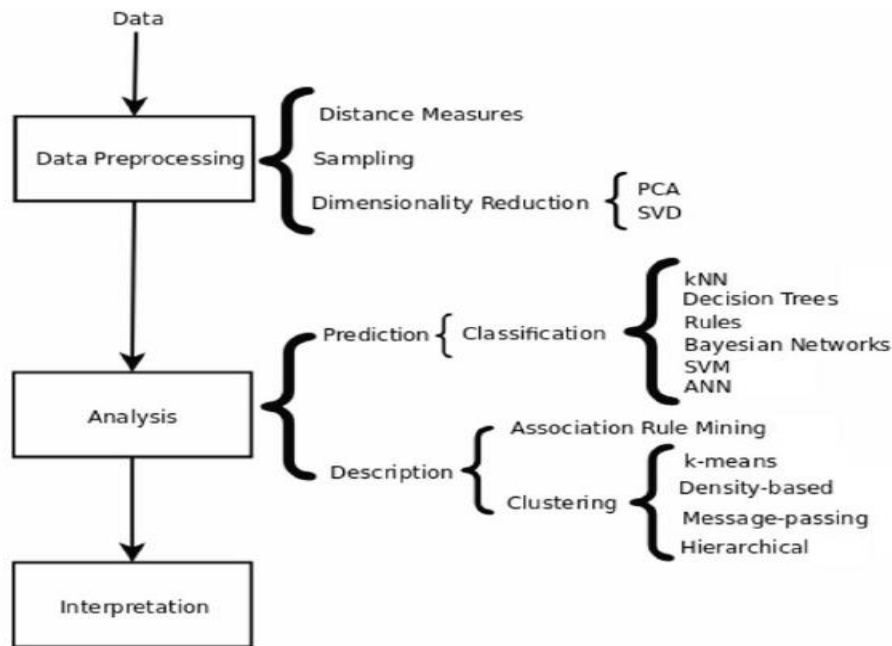
Figure 1-Main Techniques in Data Mining

Association Rules

Let $I = \{I_1, I_2, I_3, \cdots, I_m\}$ be a set of things. Give D a chance to be a set of exchange in a database where every exchange T is a set of things with the end goal that $T \subseteq I$. Every exchange in the database is related with an identifier TID, and let A be a set of things. An exchange T contains A if and only if $A \subseteq T$. An affiliation rule is a ramifications of the structure $A \implies B$, where $A \subset I$, $B \subset I$, and $A \cap B = \emptyset$. The standard $A \implies B$ holds in the arrangement of database exchanges D with support s, where s is the level of exchanges in D that contains $A \cup B$, which implies the likelihood $P(A \cup B)$ demonstrating that an exchange contains the association of set A and set B. Moreover, the certainty c of the standard $A \implies B$ in the exchange set D is the level of exchange in D that is containing A which is likewise containing B too, which implies the contingent likelihood $P(B \mid A)$. Subsequently, the guidelines that fulfil both a base support limit and a base certainty edge are called solid affiliation rules [2].

The certainty c of rule A$\implies$B can be obtained from the support tally of A and $A \cup B$ by the equation:

$$certainty(A \implies B) = P(B \mid A) = \frac{support(A \cup B)}{support(A)} = \frac{support-count(A \cup B)}{support-count(A)} \qquad (2)$$

Discovering all regular itemsets and creating solid affiliation rules are the primary procedures of affiliation rule mining. In practice, it is customary to use 35% and 60%, respectively, as minimum threshold values for support and certainty. However, this study used 50% support and 80% certainty to boost confidence in the proposed algorithm.

The Apriori Calculation

The Apriori algorithm is an algorithm for proficient affiliation rule disclosure proposed by Agrawal and Srikant in 1994 [3]. Apriori calculation utilizes a level-wise hunt strategy, where $k - itemsets$

are utilized to investigate $(k + 1) - itemsets$. A joint step is required to findL_{k-1}. A lot of applicant $k - itemsets$ can be created by joining L_{k-1} with itself and is meant C_k [4].

Contributions

The proposed algorithm utilises the Apriori algorithm on the binary appraisals lattice of user preferences tp, generating strong affiliation rules that represent clients' ratings of things in the framework's database, classified into $Favorite$ and $Non - Favorite$ items. For articles in the $Favorite$ items set, if a client has not appraised an article derived from the set in his/her preferences, the algorithm proffers such article to the client as a suggestion. On the other hand, the algorithm uses a combination of methods with the items-based approach to generate similar articles that are yet unappraised for clients by calculating the similitude between two parallel vectors representing clients' ratings and preferences. However, with a predetermined number of appraisals, the evaluation lattice $(User \times Item)$ is viewed as a scanty lattice. Executing the Apriori calculation on a scanty lattice can deliver numerous superfluous affiliation rules. The proposed algorithm avoids this unwholesome development by making several runs on the evaluation lattice until a minimum sufficient threshold number of rules are produced.

Within the available literature, the proposed referral framework is the only known system that is context-independent as it fits into more than one use-case scenario. This is due to the fact that it does not require the collection of context-aware bio data and other related statistics from users to proffer suggestions. It can thus be deployed in diverse contexts such as $(Student \times Course)$, $(Tourist \times VacationPlace)$ or $(Person \times Restaurant)$, which makes it a general-purpose utility. This is unlike the proposals of Chellatamilan and Suresh [5] and Bendakir and Aïımeur [6], as well as that of Logesh and Subramaniyaswamy [7].

The remainder of this paper is organised as follows:

Section 2 gives a brief discussion of related works on recommendation systems based on association rules mining. Section 3 presents the materials and methods of this paper and the proposed algorithm. Section 4 shows the results of the experiments of the proposed algorithm. Finally, conclusions arising from the findings of the study form the thrust of section 5.

2.Related Works

Connection rule learning is a system for finding captivating relations between factors [7], and various referral structures that use association rules mining techniques appeared in the works. Chellatamilan and Suresh [5] presented an idea for building a proposition system for the e-Learning structure using Association Rules Mining to outfit researchers with the best decision of learning materials and e-learning resources. This system used an audit review required to aggregate data from the customers. Bendakir and Aïımeur [6] proposed a course referral system reliant on connection rules. The structure merges a data mining process with customer examinations in referral.

The degree to which likeness exists between the things proposed and the clients is determined by a content-based framework [8-9]. The procedure includes the examination between the inclinations of the clients and the article highlights. The degree to which the client profiles and choices are coordinated is spoken to by a general score of execution. High execution score shows elite as for the option considered. Client's accounts are additionally considered some of the time.

Cooperative frameworks consider client clusters that have comparative likings and inclinations to make the suggestions. The client appraisals of things are utilized to decide how comparable the client's inclinations are. At the point that a set of clients is resolved with the end goal that the current client has comparative inclinations with that set, the proposals are made to the current client dependent on the inclinations of the decided set. Statistics-based frameworks utilize the statistical data of clients, for example, nationality, age and educational level, to offer recommendations. The arrangement of the stereotype run-of-the-mill classes here is one of a kind, which is unique in relation to other recommender frameworks designed for use in a general-purpose setting [10–12].

Several other multimode referral structures join at any rate two different ways to manage improved better execution and reduce the burdens of the pure referral system approaches [13, 14]. Cuts et al. [15] described the engineering of such a referral framework. Pazzani and Billsus [16] used a content-based framework in their multimode recommendation system. Their system collects data about user preferences and other feedback using the approach outlined in [17] and utilises machine learning algorithms [18].

Each of the above characterized model approaches depicts the referral framework as far as what and how the user inclinations would be in specific situations [19–20]. A strategy to circumvent the demerits of these models is to adopt a mix of more than one model. A multimode referral framework can thus be utilized to give proficient proposals to users in a general-purpose setting.

3.Materials and Methods

Figure-2 is a graphical model of the proposed multimode referral framework.

In particular, the system tends to the proposal of $Favorite$ and $Non-Favorite$ items for $Favorite$ items, the structure straightly applies the created affiliation rules to offer proposals for the client; for Non-Most loved things, the system applies a substance based way to deal with offer suggestions. The proposed calculation considers every one of the things that are evaluated by a client regardless of whether the appraisals are low. Figure-3 shows the proposed calculation.
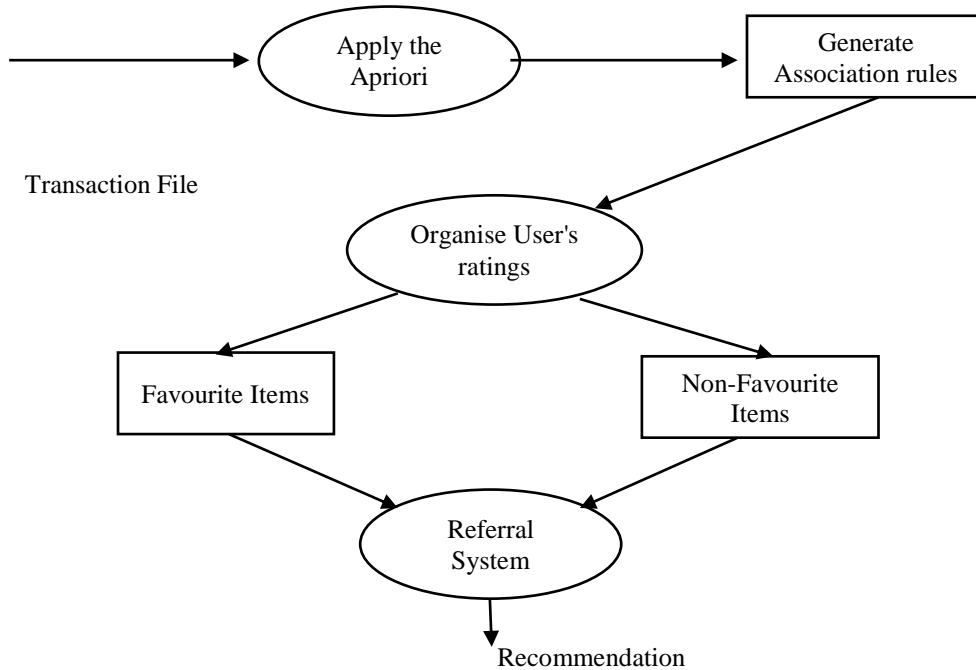
Figure 2-Graphical model of the proposed framework

Association rules are generated in the Apriori algorithm whose information sources are the exchanges record, least support, and least certainty. Table 1 is a representation of the transaction file in a matrix form.

Table 1-The transactions file in the form of a binary rating matrix

User/Item	$Item_1$	$Item_2$	$Item_3$	$Item_4$	$Item_5$	⋯	$Item_n$
$User_1$	11	1	0	0	1	⋯	1
$User_2$	0	1	0	1	1	⋯	1
$User_3$	0	0	1	0	0	⋯	0
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
$User_m$	1	0	0	0	1	⋯	0

In the above matrix, 0 means that $User_m$ has not yet ranked the $Item_n$. 1 means that $User_m$ has ranked the $Item_n$

Algorithm 1: *The Proposed Recommendation Framework*
Part I: *Generate the association rules using Apriori Algorithm*
Part II:
for *each target user, m* ***do***
find the items that the user m has ranked before
group the items that the user m has ranked into two classes:
Favourite Items Class (*rating of the items* $>= 3$)
Non − favourite Items Class (*rating of the items* < 3)
for *each item n in the Favourite Items Class* ***do***
if *the item n is in the associated items* ***then***
if *the user m has not ranked the item u derived from item n* ***then***
recommend the associated item u to the user m
end if
end if
end for
end for
for *each item k in the Non − Favorite Class* ***do***
use the Item − Based approach to find similar items for the target user m
end for
Figure 3:   **Algorithm for the proposed framework**

The Apriori algorithm generates a list of strong association rules. After this step, an item is classified as either favourite or non-favourite. Rating of the *items* > 3 implies *Favourite Items*, and rating of the *items* < 3 implies *Non − Favourite Items*. This information can be obtained from the original rating matrix, as shown in Table-2.

Table 2-The original rating matrix

User/Item	$Item_1$	$Item_2$	$Item_3$	$Item_4$	$Item_5$	$\cdots$	$Item_n$
$User_1$	51	1	$\emptyset$	3	$\emptyset$	$\cdots$	3
$User_2$	4	1	$\emptyset$	$\emptyset$	1	$\cdots$	4
$User_3$	$\emptyset$	3	$\emptyset$	4	5	$\cdots$	4
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$User_m$	$\emptyset$	4	5	2	1	$\cdots$	1

The following stage in the proposed calculation is: for each $Item_n$ in the *Favourite Items* table, check if the $Item_n$ is in the left-hand side of the created affiliation rules, and check if the user does not rate the item $Item_u$ that is in the right-hand side. Then, one can recommend the $Item_u$ to the user. The item-based approach forms the basis for the implementation of the *Non − favourite* items. The procedure is to discover things like those considered as *Non − Favourite Items* utilizing words that portray a thing as the fundamental highlights for choosing similitude among things. The similarity is represented as a vector of binary values. The proposed framework utilizes the Jaccard coefficient to gauge the closeness between two things [21]. This is utilized to process the closeness between two double vectors, and takes the following formula [22]:

$$Jaccard(i,j) = \frac{|S(i) \cap S(j)|}{|S(i) \cup S(j)|},$$ (3)

where S signifies the example set of things i and j.
Since equation (3) is used to measure the similitude between two parallel vectors, for straightforwardness, it takes the accompanying equation [23]

$$Jaccard = \frac{M_{11}}{M_{01} + M_{10} + M_{11}},$$ (4)

where M_{01} is the quantity of properties where object i was 0 and item j was 1, M_{10} is the quantity of qualities where object i was 1 and article j was 0, M_{00} is the quantity of traits where object i was 0 and article j was 0, and M_{11} is the quantity of characteristics where object i was 1 and item j was 1.

4.   Experimental Setup

The experiments used the dataset of MovieLens, given by GroupLens Research [24]. It is an open dataset consisting of 100,000 motion picture appraisals on a size of 1-5, taken from 943 clients on 1,682 motion pictures. The dataset is, as of now, tidied up with no compelling reason to preprocess the

datasets. Nonetheless, the dataset records have been reformatted to fit into the execution of the proposed calculation. WEKA software generated the association rules.

The experiment utilized a five-overlap cross-approval. At the point when the calculation creates a related motion picture for a specific client, the rating of the film is anticipated by getting the appraisals of the related motion picture from different clients that have evaluated the motion picture and then normalizing the evaluations. The exactness was estimated by utilizing two diverse assessment measurements, as described below.

Mean Absolute Error (MAE)

This is a statistical exactness metric used to gauge the normal outright deviation between an anticipated score and the user's genuine score of a thing [25]. It is a broadly utilized measurement in assessing the exactness of a proposal framework [26] and takes the structure:

$$MAE = \frac{\sum_{i=1}^{N}|p_i - r_i|}{N} \tag{5}$$

where p_i is the anticipated score, r_i is the real score, and N is the aggregate of the scores.

Root Mean Squared Error (RMSE)

This is the most well-known measurement utilized in assessing the exactness of anticipated evaluations in referral frameworks [27]. It quantifies the nature of anticipated appraisals [28] and takes the form:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(|p_i - r_i|)^2}{N}} \tag{6}$$

Experiments
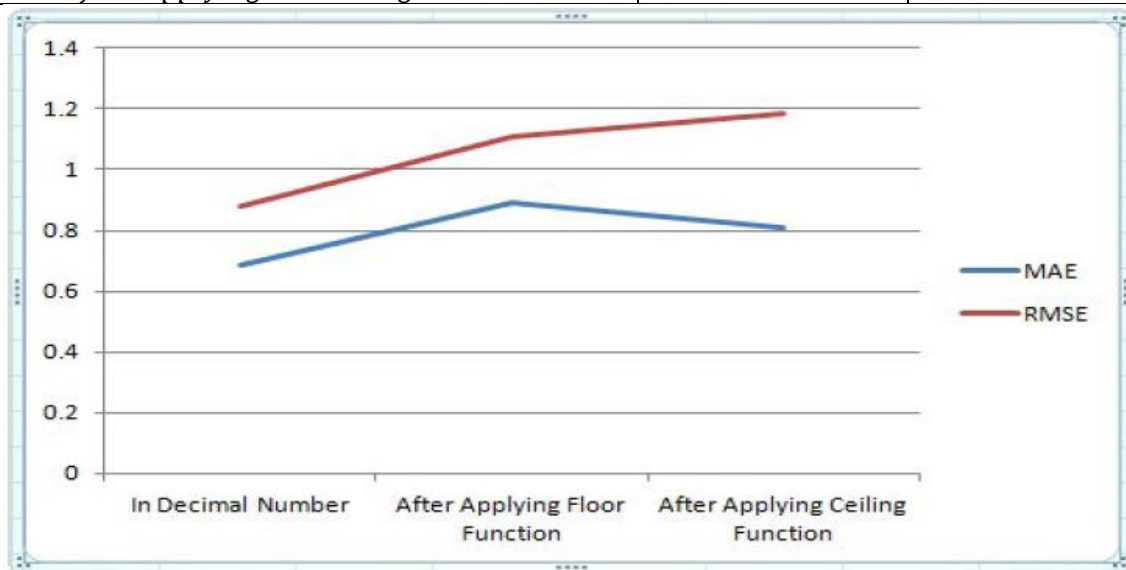
I   Favourite Item Recommendation

The transaction file for generating the association rules used the format of Table 1. A preparation dataset with things (motion pictures) that were evaluated by a minimum of 320 clients was generated. The activity delivered 12 things; the total number of clients remained at 943. WEKA produced 16 rules that were considered relevant at 35 % minimum support and 80 % confidence.

Results (Favourite Item Recommendation)

The rating matrix ($User \times Item$) with 943 clients and 12 things (that have been evaluated by a minimum of 320 clients) was utilized in this trial for each of the five-overlap cross-approvals in WEKA. The results were evaluated in three different cases and are summarised in Table-3 and Figure-4.

Table 3-Evaluation of the experiment's results

Evaluation	*MAE*	*RMSE*
In Decimal Numbers	0.6865502804	0.8786227744
After Applying the Floor Function	0.892116171	1.10688302
After Applying the Ceiling Function	0.806277217	1.86032072

Figure 4-The results of the evaluation for the three cases

From Table-3 and Figure-4, it is apparent that the anticipated appraisals in denary give a better anticipated rating.

II   Non-Favourite Items Recommendation

To actualize the second piece of the proposed system, the traits that portray the thing were considered. Every film is defined by its class in binary values and represented as a vector, as shown in Table 4.

Table 4-The representation of a movie

Movies/Genres	Action	Adventure	Animation	⋯	Western
$Movie_1$	0	1	0	⋯	0
$Movie_2$	1	0	0	⋯	1
$Movie_3$	0	1	0	⋯	1
$Movie_4$	1	1	0	⋯	0
⋯	⋯	⋯	⋯	⋯	⋯
$Movie_n$	1	0	1	⋯	0

Equation (4) is then applied to gauge the similitude between the film that was not liked (in the $Non-Favourite\ Items$ classification) by a client and different motion pictures that were not seen at this point, and returns most comparative motion pictures to the client. Table 5 gives the outline of the consequences of the assessment of the $Non-Favourite\ Items$ part in the context of Equation (4) with a similitude of 50 % or more among the motion pictures.

Table 5-Summary of the experiment's results of $Non-Favourite$ items

Evaluation	MAE	RMSE
In Decimal Numbers	0:871955461	1:104979437
After Applying the Floor Function	1:161398927	1:474234682
After Applying the Ceiling Function	1:304981189	1:632056341
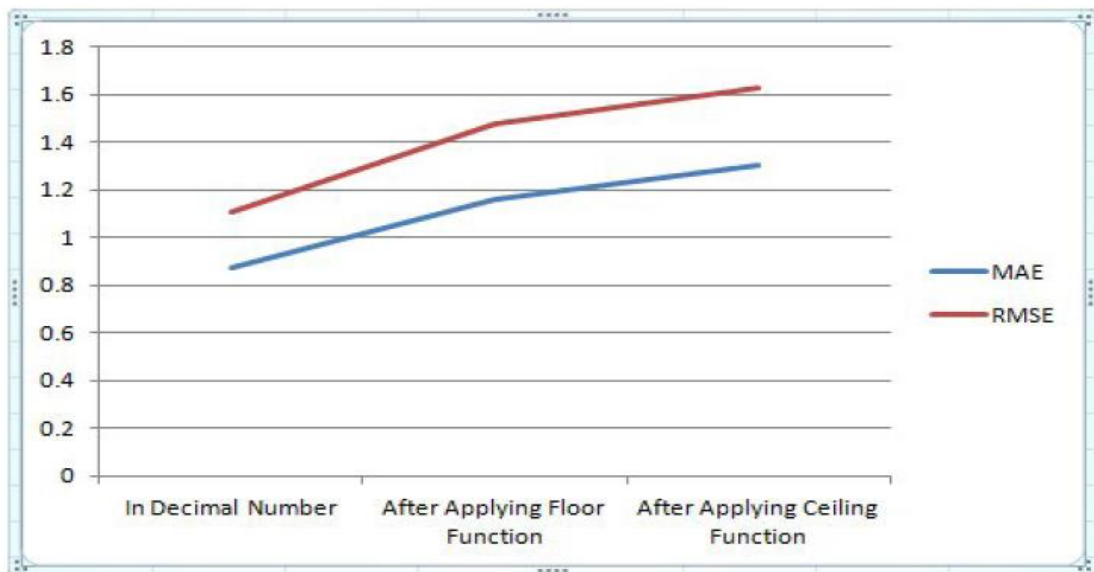
Figure 5-Experiment's results of $Non-Favourite$ items

The outcomes of the analysis on $Non-Favourite\ Items$ from Table-5 and Figure-5 show that the anticipated appraisals in denary gives more exact anticipated evaluations than the other floor and ceiling utilities.

Discussion

The main problems in the design of referral systems are versatility and sparsity. In the proposed framework, bunching and similitude prediction techniques are utilized to overcome these issues. Also, affiliation rule mining and article-based data were additionally utilised to overcome the cold start issue, consequently expanding the precision of the recommendation. To assess the model, a huge scale datasets of MovieLens [24] was used. The outcomes of the proposed framework demonstrated that the

utilization of logical data, with the assistance of bunching, similitude calculation and affiliation rule mining, are effective in improving the efficiency of the proposed framework.

With respect to versatility, the proposed model improved the versatility of the recommendation through the utilization of bunching and the likeness forecast strategy, and the outcome is considerably better than those obtained using different techniques [5, 6]. Concerning sparsity, the proposed framework outperformed the baseline approaches [7]. Also, the model uses an affiliation rule mining method for better forecast exactness that was contrasted with that applied by another published models [12]. The improvement in the exactness of the proposed framework is a result of the combination of recommendation approaches used to aggregate user appraisals and preferences. This makes the framework deployable in diverse contexts such as ($Student \times Course$), ($Tourist \times VacationPlace$) or ($Person \times Restaurant$), thus making it versatile in a general-purpose setting.

5. Conclusions

This study proposed a multimode referral framework to be applied on a two- dimensional ($User \times Item$)-space with an enormous client base and relatively few offerings. The proposed structure utilizes both $Favourite\ Items$ and $Non-Favourite\ Items$ of a specific customer, predicated on the reconciliation of affiliation rules mining and the substance-based methodology. With a predetermined number of appraisals, in any case, the evaluation lattice ($User \times Item$) is viewed as a scanty lattice; executing the Apriori calculation on a scanty lattice can deliver numerous superfluous affiliation rules. It is thus beneficial to find a specific method to handle the scantiness from the evaluation lattice.

References

1. Adomavicius, G. and Tuzhilin, A. **2005**. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering,* IEEE Transactions on, **17**(6): 734-749.
2. Han, J. and Kamber, M. **2006**. Data mining: concepts and techniques 2nd ed. Amsterdam: Elsevier.
3. Hegland, M. **2007**. The Apriori Algorithm Tutorial: Mathematics and Computation in Imaging; *Science and Information Processing,* **11**: 209-262.
4. Han, J., J., Pei, and Yin, Y. **2000**. Mining Frequent Patterns without Candidate Generation. In ACM SIGMOD Record; **29**(2): 1-12, ACM.
5. Chellatamilan T. and Suresh R. **2011**. An e-Learning Recommendation System using Association Rule Mining Technique; *European Journal of Scientific Research*, **64**(2): 330-339.
6. Bendakir, N. and Aiimeur E. **2006**. Using Association Rules for Course Recommendation. In Proceedings of the AAAI Workshop on Educational Data Mining; pp. 31-40.
7. Logesh R. and Subramaniyaswamy, V. **2019**. "Exploring Hybrid Recommender Systems for Personalised Travel Applications." Advances in Intelligent Systems and Computing, 768 434-443.
8. Liu B., Hsu W., and Ma Y. **1998**. Integrating Classification and Association rule mining. In Proceedings of the 4th international conference on Knowledge Discovery and Data mining KDD'98, pp. 80–86.
9. Zhang Y., Tu, Z. and. Wang, Q. **2017**. "Temporal-Topic Based Recommender for Social Network Services," *Mobile Networks and Applications,* **22**: 1182–1191.
10. Soni K., Goyal R., Vadera B. and More S. **2017**. "A Three Way Hybrid Movie Recommendation System," *International Journal of Computer Applications*, **160**(9): 29–32.
11. Subramaniyaswamy V. and Logesh R. **2017**. "Adaptive KNN based recommender system through mining of user preferences." *Wireless Pers. Commun.* **97**(2): 2229–2247.
12. Wang Y. and Xu W. **2018.** "Leveraging deep learningwith LDA-based text analytics to detect automobile insurance fraud," *Decision Support Systems*, **105**: 87–95,
13. Ghazanfar, M.A. and Prugel-Bennett A. **2010**. A scalable, accurate hybrid recommender system. In Knowledge Discovery and Data Mining WKDD'10. Third International Conference on AI; pp. 94-98, IEEE.
14. Burke, R. **2002**. Hybrid recommender systems: Survey and experiments. *User modelling and user-adapted interaction,* **12**(4): 331-370.

15. Lops P., Gemmis M. and Semeraro, G. **2011**. *Content-based recommender systems: State of the art and trends*. Recommender Systems Handbook, pp.73-105

16. Pazzani M. J. and Billsus D. **2007**. Content-based Recommendation Systems. *Adaptive Web: Methods and Strategies of Web Personalisation,* **4321:** 325–341. Springer-Verlag.

17. Rocchio J. J. **1971**. Relevance feedback in information retrieval. In G. Salton, (ed.) *The Smart retrieval system – experiments in automatic document processing*. Englewood Cliffs, NJ; pp. 313–323.: Prentice-Hall,

18. Han J., Kamber M. and Pei J. **2011.** *Data Mining: Concepts and Techniques*. 3rd ed. San Francisco, CA., USA. Morgan Kaufmann Publishers Inc.

19. Zhang Y., Chen M., Huang D., Wu, D. and Li Y. **2017**. "IDoctor: personalized and professionalized medical recommendations based on hybrid matrix factorization," *Future Generation Computer Systems*, **66**: 30–35.

20. Nilashi, M., Ibrahim, O. and Bagherifard, K. **2018**. "A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques." *Expert Syst. Appl*. **92**: 507–520.

21. da Silva Meyer, Garcia A. F., de Souza A. P. and de Souza C. L. **2004**. Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L.). *Genetics and Molecular Biology,* **27**: 83-91.

22. Sigurbjrnsson, B. and Van Zwol R. **2008**. Flickr tag recommendation based on collective knowledge. In Proceedings of the 17th International Conference on World Wide Web; pp. 327-336. ACM.

23. Tan, P., Steinbach M. and Kumar V. **2005**. *Introduction to Data Mining*; Boston: Pearson Addison Wesley.

24. MovieLens Datasets. **2011**. *GroupLens Research*. Retrieved September 18, 2019, from http://www.grouplens.org/node/73

25. Herlocker, J. L., Konstan J. A., Terveen L. G. and Riedl J. T. **2004**. Evaluating Collaborative Filtering Recommender Systems; *ACM Transactions on Information Systems* (TOIS), **22**(1): 5-53.

26. Sarwar, B., G., Karypis, J., Konstan, and Riedl J. **2001**. Item-based Collaborative Filtering Recommendation Algorithms. In Proceedings of the 10th International Conference on World Wide Web; pp. 285-295. ACM.

27. Shani, G. and Gunawardana A. **2011**. *Evaluating Recommendation Systems*. Handbook of Recommender Systems, pp. 257-297.

28. Koren, Y. **2008**. Factorisation meets the neighbourhood: A Multifaceted Collaborative Filtering Model. In Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; pp. 426-434. ACM.