Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING COLLABORATIVE FILTERING: ADDRESSING SPARSITY AND GRAY SHEEP WITH OPPOSITE USER INFERENCE

ABDELLAH EL FAZZIKI[1], YASSER EL MADANI EL ALAMI[2], JALIL ELHASSOUNI[2], MOHAMMED BENBRAHIM[1]

[1]LIMAS, University of Sidi Mohamed Ben Abdellah, Fez, Morocco.

[2]Mohammed V University, Rabat, Morocco

Email: [1] elf.abdellah@gmail.com , mohammed.benbrahim@usmba.ac.ma,
[2] y.alami@um5s.net.ma, jalil.elhassouni@gmail.com

ABSTRACT

Collaborative filtering (CF) is a popular recommendation approach which seeks to find similar users to predict what an active user might like. However, CF suffers from two main challenges: sparsity and gray sheep. In both cases, recommending useful items is a difficult task. In this paper, we propose a new approach to address these challenges. It consists of combining Singular Value Decomposition and Association rule methods with enriched rating matrix. In addition to actual users, this matrix incorporates virtual users inferred from opposing ratings provided by real users. Our approach attempts to increase the density of similar users and makes it easier to make useful recommendations. We conducted a comparative study showing that our method outperformed traditional CF approaches in terms of accuracy.

Keywords: *Recommendation System; Collaborative Filtering; Opposite Preferences; Model-Based CF; SVD; Association Rules; Sparsity Problem.*

1. INTRODUCTION

In recent years, the utilization of recommendation systems has surged, particularly in ecommerce and online media. Many companies have adopted these systems to enhance the sales by suggesting or recommending additional products and services to their users. As a result, these systems have transformed how online businesses interact with both existing and potential customers. The transition from physical stores to the digital sphere has forced many companies to adjust their strategies to align with consumer demands [1]. In ecommerce, the primary objective of recommendation systems is to guide customers in making informed purchase choices and down the long list of products to a more personalized one. This can manifest as product suggestions or product reviews on retail websites [1] such as Amazon [2], Netflix [3], and Spotify [4][5]. These platforms have enjoyed tremendous success by making entertainment and shopping more accessible to consumers and offering an incredible experience, especially during the Covid19 pandemic.

To our best knowledge, there are three main types of recommender systems: Content-based, collaborative, and hybrid. Although they are efficient and simple, the collaborative filtering (CF) approach remains the most widely used in recommendation systems [6].

The main assumption behind CF is based on users or items with similar interests and preferences. These preferences can be communicated in a variety of ways, including explicit feedback (i.e., ratings, likes, etc.) [7], or implicit feedback inferred from users' behavior (purchases' history, time spent on web content, etc.) [8]. They can be presented as a matrix known as a rating matrix [9].

Two types of approaches for collaborative filtering have been proposed: memory-based CF and model-based CF. In Memory-based CF, recommendations are generated based on the whole rating matrix [10][11][12]. Besides, Model-based approach uses machine learning algorithms to build the model which will generate suitable items for each user in a personalized manner[13]. This approach uses machine learning techniques such as Association Rule (AR), clustering techniques [14], Matrix factorization (Singular Value Decomposition (SVD), Funk Singular Value Decomposition (Funk SVD)), etc. Despite the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

improvement made in collaborative filtering methods, they still encounter many challenges that limit the quality of recommendations, notably the gray sheep and sparsity problems. The sparsity refers to the issue of having a large amount of data with very few ratings or interactions[15], while the gray sheep refers to a category of users who exhibit preferences or behaviors that deviate significantly from the broader user population[16].

In this work, we propose a new approach which consists of combining SVD and Association rule methods with enriched rating matrix.

This paper focuses on discussing machine learning methods used in collaborative filtering that is widely used today in this field: SVM, SVMF and AR. We list recommender systems' issues and challenges and how to overcome these challenges; finally, we present our approach that surpasses many traditional recommendation approaches and solves sparsity and gray sheep problems.

The remainder of the paper is organized as follows: in the next section, we present the traditional CF methods followed by an overview of the related work concerning the challenges of gray sheep and data sparsity in recommendation systems in section 3. In section 4, we outline our approach designed to address the challenges of data sparsity and gray sheep. Then, we present our experimental analysis, which includes evaluation metrics and various experimental results in section 5. Finally, in Section 6, we draw conclusions and suggest potential directions for future research.

2. RELATED WORK

In this section, we discuss the issues of sparsity and gray-sheep users, addressing a gap in current research on recommender systems. We also discuss the subject of compelling opposite user ratings, and how this idea is exploited in recommender systems research. Thereafter, we mention some work that examines the CF with the technique of SVD and AR.

According to Claypool et al. [17], the effectiveness of the traditional CF algorithm is not consistent among users. Users can be categorized into two primary groups: White Sheep (WS) and Gray Sheep (GS). WS users exhibit high similarity to many other users, indicated by a high correlation coefficient. However, GS users are characterized by their dissimilarity or partial similarity to other users, resulting in lower correlation coefficients with most users [18]. As a consequence, recommendations for GS users tend to be less accurate [19], and, thus, these users do not fully benefit from traditional recommendation systems. While some efforts have been made to address the issue of GS users [20][21], Claypool et al. [17] emphasized the problem of GS users and introduced a hybrid recommendation system aimed at providing updated recommendations. This hybrid system combines both Collaborative Filtering (CF) and content-based filtering approaches through an average weighted method. However, it is worth noting that their approach did not specifically target GS users, nor did it offer a formal solution for this particular problem [22]. Furthermore, their method was tested using the MovieLens dataset in a CF domain, but it was primarily a simulation. They did not outline a method for identifying GS users or catering to their unique needs. One potential approach to identifying GS users is through the utilization of clustering algorithms offline. By employing clustering methods, it is possible to identify GS users based on empirical similarity thresholds, effectively separating them from the remaining user clusters [23]. The issue of data sparsity arises when the rating matrix, which contains essential details about the ratings, is notably sparse. This sparsity can lead to inefficiencies within recommender systems. This problem is further divided into two categories: limited coverage and transitivity [24]. The data sparsity is a problem that surrounds RS in general and CF in particular. To overcome this problem, various approaches and methods have been offered. The majority of them receive positive feedback. However, a general solution for dealing with sparsity is still elusive [25][26][27][28].

Furthermore, Association Rule (AR) algorithms are instrumental in establishing relationships between items within a transaction. These rules, such as A --> B, are applicable in recommendation models to anticipate user and item characteristics [25][26]. The association rules-based data mining strategy is the one that is being studied the most [29]. AR specifies the rules about how one event is related to another. It is also a sort of clustering that categorizes data based on its significance [30]. When Event A occurs, an association rule is given in the form R: A –> B, indicating that Event B may occur as well. This combination can reduce the size and processing time of massive data.

In the context of recommendation systems, Singular Value Decomposition (SVD) is a well-established method employed to address challenges faced by collaborative filtering[31]. SVD is a popular approach for recommendation filtering that decomposes an m x n matrix A into three matrices utilizing the formula $A = USV^T$ [31].

Funk SVD, a variant of Regularized SVD, aims to create a matrix decomposition that closely approximates the real matrix's values with minimal error[32]. It is noteworthy that Funk SVD achieved success in winning the Netflix recommendation competition [33]. Funk SVD is a specialized form of SVD designed for recommendation systems. Funk SVD incorporates optimization techniques and is tailored to handle the unique issues associated with recommendation data, making it more effective in this context. Simon Funk [33] suggested regularized SVD called Funk SVD for collaborative filtering, a methodology inspired by effective methods from the domain of natural language processing [34]. The proposal of learning rate and regularization constants, as well as a method of clipping predictions, are included in Simon Funk's description [35]. Funk Singular Value Decomposition model (FSVD) is a variant of Matrix Factorization (MF) introduced by Funk in 2006. It gained prominence for its superior performance in the Netflix Prize competition, which aimed to improve recommendation algorithms for movie ratings. The central concept behind FSVD, as in traditional MF, is to represent users and items using latent feature vectors derived from the rating matrix. When there is a strong correspondence between the feature vectors of users and items, it facilitates making accurate recommendations [36].

Applying dimension reduction methods to the rating matrix is the most popular methodology [37][38]. The sparsity problem is addressed by the dimension reduction method, which condenses the matrix by deleting low representative or noisy data.

Other methods, such as information retrieval's and Latent Semantic Indexing (LSI), are frequently used [39]. One disadvantage of these approaches is that during the reduction process, potentially beneficial information may be inevitably deleted. Utilizing associative retrieval techniques toward the bipartite graph consisting of items and users [40], content boosted CF [41], singular value decomposition techniques (SVD)[42] and the leverage of item-based similarity to replace user-based similarity [43] are some additional methods to address sparsity.

These methods build a model that learns from user-item interactions using low dimensional representations as a component (user and item feature vectors) [44]. The SVD model includes imaginary users' data to enhance rating matrix in the matrix factorization process, resulting in better recommendations than traditional SVD.

The works of [45][46] confirm that the efficiency of the traditional CF algorithm is less efficient than the approach which is based on the increase of the rating matrix. The latter is done by the utilization of imaginary users being the opposite of real users. The same idea is used in [47] and confirms that this idea improves the prediction and solves the gray sheep problem. Yet, both works use the memory-based CF approach.

There are several overall technical gaps that are observed in the existing works that led to the design of the proposed approach. First, the available recommendation systems do not focus on efficiency or achieving the ultimate goals for each recommendation system, especially in movie recommendation.

Nonetheless, it is imperative to note that the approaches mentioned above do not provide a precise solution for the issues of data sparsity and gray-sheep users, and tend to ignore these problems. Consequently, when issues like data sparsity and gray-sheep users are present, recommender systems do not find sufficient data to provide effective recommendations. Therefore, in this study, we present our approach to address these problems.

3. OUR APPROACH

As stated earlier, the basic CF approach (traditional approach) depends just on rating database whether to predict similarity between user/item or to train a model. Because most of the items are only rated by a few users one of the reasons for sparsity and gray-sheep users and the resulting user-item matrix is relatively due to a lack of rating data, making algorithms ineffective.

Our research is built upon the foundation laid by [45]. The fundamental idea underpinning our approach revolves around the incorporation of an extra phase into the traditional collaborative filtering process, which we term "rating matrix enriched". It is noteworthy that the work presented in[45] predominantly centers around a memory-based approach. In contrast, our study adopts a model-based approach, leveraging techniques like AR, SVD, and FSVD to advance and broaden the matrix rating augmentation concept.

Our revised approach introduces an extra phase called "ratings matrix enriched" [45](Figure 1), aimed at expanding the rating matrix.

Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 1: The new process of CF approach

Figure 1 illustrates the proposed new process of CF approach with the addition of a new phase called rating matrix enriched.

Our enriched method is founded on the concept that consumers must share similar interests. If user X's interests are completely contrary to those of user Y, the imaginary user ¬X would have those interests. As a consequence, more data will be sent to the recommendation engine in order to suggest potential and appropriate recommendations.

The enrichment step of the ratings matrix requires adding rows to the matrix that represent opposing users to real users. The imaginary user is created by calculating the opposite preference from each item evaluated using the formula:

$$\neg \boldsymbol{r_{uj}} = \boldsymbol{Max} - \boldsymbol{r_{uj}} + \boldsymbol{Min} \quad (1)$$

$\boldsymbol{r_{uj}}$: the rating of user $\boldsymbol{u}$ for an item $\boldsymbol{j}$.

Max and ***Min***: the high and low values respectively in a given numerical scale.

Example:

If a user $\boldsymbol{u}$ rates an item with a rating of $r_{uj}= 5$, the anticipated rating of the user u will be $\neg \boldsymbol{r_{uj}}= \boldsymbol{1}$. If a user u rates an item with a rating of $\boldsymbol{r_{uj}}= \boldsymbol{2}$, the anticipated rating of the user u will be $\neg \boldsymbol{r_{uj}}= \boldsymbol{4}$.

We enhance the density of the rating matrix by incorporating opposite user preferences. This enriched matrix is then utilized in combination with SVD and association rule techniques to effectively tackle the issues of data sparsity and gray-sheep users.

4. EXPERIMENTAL RESULTS

In this section, we describe the dataset performance measurement and results for the comparison between our proposed CF approach and traditional CF approaches.

In this section, several tests have been carried out to demonstrate the originality and efficacy of our approach. As a result, we divide our dataset into two parts: 80 percent for training and 20 percent for testing (Figure 2); this concept is known as Train/Test Split.

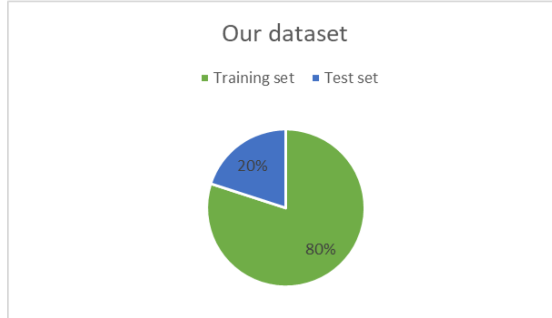

Figure 2: Divided dataset

Utilizing the Recommenderlab package [37] and the MovieLens dataset, we successfully developed a film recommendation system in R.

The goal is to compare the performance of our suggested approach to that of traditional approaches using realworld datasets. The assessment approach and the specification test environment will be followed by a brief discussion of the datasets used. Therefore, the results were derived from comparisons in order to determine the most effective approach.

4.1 Datasets collection

The MovieLens dataset is used to assess the performance of the suggested technique.

The University of Minnesota's GroupLens Research Project [48] generated MovieLens datasets, and MovieLens is a webbased research recommender system that was originally released in Fall 1997; the data is freely accessible via: https://grouplens.org/datasets/movielens/.

Every week, hundreds of people visit MovieLens to rate and receive movie recommendations. The ratings are on a five point scale, with 1 and 2 indicating negative ratings, 4 and 5 indicating positive ratings, and 3 indicating ambivalence [27]. It includes 1682 films, 943 users, and 100,000 rankings.

4.2 Experiments

The experimental evaluation of our suggested method is carried out in this section, and the results are based on a variety of frequently used metrics with various parameters.

All of the algorithms were written in the R programming language and ran on an Intel i7 2.4 GHz processor with 8 GB of RAM running Windows 8.1, using MovieLens datasets. The choice of R for constructing the film recommendation system is justified due to its robust statistical capabilities, specialized machine learning libraries like RecommenderLab, and strong support within the data science community, providing an efficient environment for developing and prototyping recommendation algorithms.

4.3 Performance measurement

We apply statistical accuracy measurements in this paper [23]. Statistical accuracy metrics compare the numerical deviation of the predicted ratings from the actual user ratings to determine the accuracy of a prediction algorithm. In the context of evaluating recommendation systems, metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE) are frequently employed. These metrics provide a quantitative measure of the accuracy and precision of a recommendation system. MAE calculates the average absolute differences between predicted and actual ratings, RMSE measures the square root of the average squared differences, and MSE represents the average squared differences between predicted and actual ratings. These metrics play a crucial role in assessing the performance and effectiveness of recommendation algorithms. All of the aforementioned measures were calculated using the same data (MovieLens datasets), and they produced similar results.

MAE, MSE and RMSE measurements calculate the differences between our results and this facetious reality.

The goal in this section is to evaluate MAE, MSE and RMSE for each set of observations.

Tableau 1: comparison of RMSE, MSE and MAE between AR and enriched_AR for 10 folds

Testing set		RMSE	MSE	MAE
Fold 1	AR	1,852	2,032	1,342
	Enriched_AR	1,502	1,82	1,025
Fold 2	AR	1,822	2,056	1,369
	Enriched_AR	1,523	1,798	1,009
Fold 3	AR	1,789	2,112	1,379
	Enriched_AR	1,498	1,852	1,011
Fold 4	AR	1,955	2,089	1,396
	Enriched_AR	1,598	1,85	1,023
Fold 5	AR	2,002	2,189	1,401
	Enriched_AR	1,6	1,935	1,156
Fold 6	AR	1,785	2,006	1,299
	Enriched_AR	1,598	1,796	1,123
Fold 7	AR	1,986	2,103	1,388
	Enriched_AR	1,536	1,758	1,099
Fold 8	AR	1,896	2,156	1,379
	Enriched_AR	1,489	1,832	1,008
Fold 9	AR	1,763	2,098	1,399
	Enriched_AR	1,423	1,864	1,096
Fold 10	AR	1,754	2,179	1,389
	Enriched_AR	1,453	1,897	1,097
Average	***AR***	***1,8604***	***2,102***	***1,3741***
	Enriched_AR	***1,522***	***1,8402***	***1,0647***

Table 1 presents a comparative analysis of the outcomes from 10 experiments conducted using AR (Association Rules) and enriched AR (Enriched Association Rules). The evaluation metrics employed for comparison are RMSE, MSE, and MAE.

The findings demonstrate that the RMSE, MSE, and MAE values obtained using the enriched AR approach are consistently lower than those obtained using AR. This lower value across all three evaluation metrics collectively leads to a reduction in the average results for enriched AR compared to those of AR. This indicates that the enriched AR approach outperforms the standard AR approach in terms of predictive accuracy. The figure below illustrates the average results from ten

Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

experiments, providing a comparison between enriched AR and traditional AR.

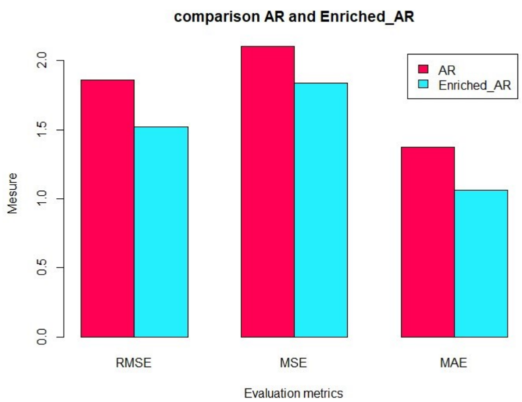

Figure 3: Comparison of AR and Enriched_AR

Figure 3 provides a comprehensive comparison between our proposed approach, Enriched Association Rule (enriched_AR), and the traditional method, Association Rule (AR). The bar graph provides a visual representation of the comparison between these two approaches based on the average evaluation metrics RMSE, MSE, and MAE.

In the first scenario, represented in Figure 3, the RMSE values show that AR scores 1.86, while enriched_AR performs better with a lower value of 1.52. In the second case, focusing on MSE, AR achieves 2.1, whereas enriched_AR attains a lower score of 1.84. Lastly, in the third case, examining the MAE metric, AR yields 1.37, while enriched_AR excels with a notably lower score of 1.06. These results consistently indicate that enriched_AR outperforms AR in all three scenarios.

The graphical representation of these results in the bar graph clearly illustrates that, across all evaluation metrics (RMSE, MSE, and MAE), AR consistently yields higher values compared to enriched_AR. This provides strong evidence that our proposed approach, enriched_AR, is superior to the traditional approach, AR.

Tableau 2: Comparison Of RMSE, MSE And MAE Between SVD And Enriched_SVD For 10 Folds

Testing set		RMSE	MSE	MAE
Fold 1	SVD	1,21011	1,25655	0,89236
	Enriched_SVD	0,7435	0,94162	0,47278
Fold 2	SVD	1,24478	1,2775	0,94858
	Enriched_SVD	0,74024	0,98497	0,49057
Fold 3	SVD	1,21982	1,23805	0,9204
	Enriched_SVD	0,73189	0,92665	0,44087
Fold 4	SVD	1,20832	1,23142	0,95088
	Enriched_SVD	0,74439	0,96595	0,45655
Fold 5	SVD	1,20244	1,25993	0,90302
	Enriched_SVD	0,71236	0,98328	0,41388
Fold 6	SVD	1,2325	1,23727	0,9634
	Enriched_SVD	0,72077	0,93343	0,40325
Fold 7	SVD	1,20086	1,2997	0,94676
	Enriched_SVD	0,71125	0,95438	0,4095
Fold 8	SVD	1,22357	1,28151	0,93541
	Enriched_SVD	0,74187	0,95739	0,46279
Fold 9	SVD	1,21788	1,2579	0,9028
	Enriched_SVD	0,72557	0,97042	0,43356
Fold 10	SVD	1,23759	1,2579	0,94779
	Enriched_SVD	0,70121	0,99907	0,42491
Average	***SVD***	***1,219787***	***1,259773***	***0,93114***
	Enriched_SVD	***0,727305***	***0,961716***	***0,440866***

Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2 offers a comprehensive analysis of the results obtained from 10 experiments comparing SVD and enriched SVD. The evaluation metrics deployed in this analysis are RMSE, MSE, and MAE.

The results consistently reveal that the values of RMSE, MSE, and MAE obtained using the enriched SVD approach are consistently lower than those obtained using the standard SVD. This consistent reduction in values across all three evaluation metrics contributes to an overall decrease in the average results for the enriched SVD approach compared to the standard SVD. This implies that the enriched SVD approach surpasses the conventional SVD approach in terms of predictive accuracy. The following figure depicts the average outcomes from ten experiments, presenting a comparison between enriched SVD and traditional SVD.

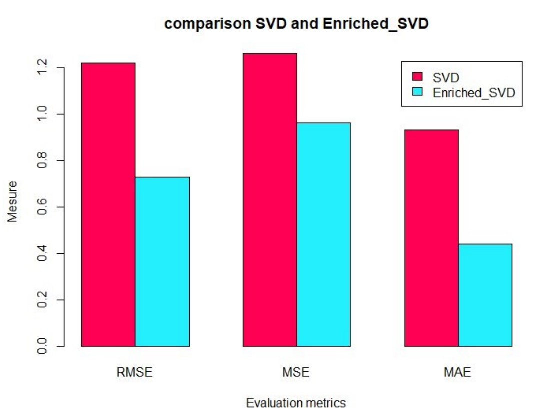

Figure 4: Comparison of SVD and Enriched_SVD method

In Figure 4, we present a comparative analysis of the average RMSE, MSE, and MAE measures between our two approaches: SVD represented by the red bars, and enriched_SVD represented by the blue bars. The x-axis of the graph displays the evaluation metrics RMSE, MSE, and MAE, while the y-axis represents a measurement scale ranging from 0.0 to 2.0.

For RMSE, the red bar, corresponding to SVD, indicates a value of 1.21 on the average measurement scale along the x-axis. In contrast, the blue bar, representing enriched_SVD, shows a significantly lower value of 0.72, indicating superior performance compared to SVD in terms of RMSE.

Similarly, when evaluating MSE, the red bar for SVD displays a value of 1.25 on the x-axis. Meanwhile, the blue bar for enriched_SVD demonstrates a lower value of 0.96, further highlighting the enhanced performance of enriched_SVD in comparison to SVD with respect to MSE.

Lastly, examining MAE, the red bar, denoting SVD, indicates a value of 0.93 on the x-axis. In contrast, the blue bar, representing enriched_SVD, reveals a substantially reduced value of 0.44, underscoring the improved performance of enriched_SVD in relation to MAE when compared to SVD. Hence, the results of this graph show that our approach (enriched_SVD) is more effective and performs better than the traditional approach (SVD) in all (RMSE, MSE, and MAE) cases.

Tableau 3: comparison of RMSE, MSE and MAE between FSVD and enriched_FSVD for 10 folds

Testing set		RMSE	MSE	MAE
fold1	FSVD	1,0995458	1,2090009	0,8685648
	Enriched_FSVD	0,94073	0,94762	0,8656485
fold 2	FSVD	1,085267	1,187536	1,084247
	Enriched_FSVD	0,91999	0,81466	0,64318
fold 3	FSVD	1,104534	1,18234	0,90273
	Enriched_FSVD	0,82788	0,89096	0,70782
fold 4	FSVD	1,1000983	1,07594	0,99912
	Enriched_FSVD	0,87149	0,99031	0,74489
fold 5	FSVD	1,1056543	1,12792	0,94007
	Enriched_FSVD	0,95691	0,93999	0,67979
fold 6	FSVD	1,1007852	1,02487	0,86891
	Enriched_FSVD	0,81719	0,84763	0,60187
fold 7	FSVD	1,0827826	1,14485	0,94529

Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	Enriched_FSVD	0,87548	0,99797	0,67487
fold 8	FSVD	1,1209832	1,13386	0,84648
	Enriched_FSVD	0,95354	0,95324	0,71967
fold 9	FSVD	1,084165	1,09896	0,91404
	Enriched_FSVD	0,86949	0,86621	0,69407
fold 10	FSVD	1,0651261	1,10651	0,82199
	Enriched_FSVD	0,83651	0,88177	0,62448
Average	**FSVD**	**1,09489415**	**1,12917869**	**0,91914418**
	Enriched_FSVD	**0,886921**	**0,913036**	**0,69562885**

Table 3 provides a thorough analysis of the outcomes from 10 experiments where FSVD and enriched FSVD were compared. The results consistently demonstrate that the values of RMSE, MSE, and MAE achieved using the enriched FSVD approach are notably lower than those obtained with the standard FSVD. This consistent reduction in values across all three evaluation metrics contributes to an overall decrease in the average results for the enriched FSVD approach in comparison to the standard FSVD. This suggests that the enriched FSVD approach excels over the conventional FSVD approach in terms of predictive accuracy.

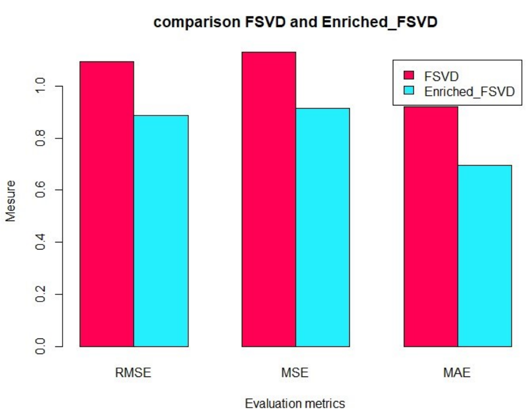

Figure 5: Comparison of FSVD and Enriched_FSVD method

Figure 5 presents a visual comparison of the average evaluation metrics RMSE, MSE, and MAE for FSVD and enriched_FSVD. These metrics are represented on the x-axis, while the measurement scale is on the y-axis. In each of the three cases (RMSE, MSE, and MAE), the bar representing FSVD is consistently higher than the bar for enriched_FSVD. This consistent pattern across all three metrics leads to the conclusion that our approach, enriched_FSVD, outperforms the traditional approach (FSVD) under all circumstances.

All figures show that our approach has a lower MAE, RMSE and MSE than the traditional approach for the MovieLens dataset.

4.4 Statistical inference

In most experiments, ensuring that the observed difference between the proposed method and the baseline is statistically significant is of the utmost importance. This is vital to confirm that the disparity is not merely a result of chance or random noise in the data. The most suitable statistical test for such comparisons is the t test (student test). The t test is a nonparametric statistical method, meaning it does not rely on specific data distribution assumptions. Its primary objective is to determine whether the population distributions are the same or different. In this context, our null hypothesis posits that the results obtained from enriched algorithm and traditional algorithm from identical populations, thereby ruling out any small gains or losses due to random chance.

Our null hypothesis posits that the outcomes derived from the enriched algorithms and those from the traditional algorithms belong to identical populations, signifying that any marginal improvements or deteriorations observed are statistically insignificant. As a customary practice, we reject the null hypothesis when the pvalue falls below a predefined threshold, which is often set at 0.05. In simpler terms, if the pvalue is less than 0.05, we can reasonably conclude that the disparities are statistically significant.

Therefore, we conducted a comparison of pvalues between RMSE, MSE, and MAE using the ttest, and the results are presented in Table 4.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Tableau 4: Comparison of pvalue between RMSE, MSE and MAE by t test for enriched algorithms and traditional algorithms

	RMSE	MSE	MAE
EnrichedAR/AR	8,11096455E08	8,39879E09	4,63195E08
EnrichedSVD/ SVD	4,2122E14	3,31869E11	1,54672E11
EnrichedFSVD/ FSVD	1,23187E07	6,71672E06	2,88269E06

As indicated in Table 4, all the computed pvalues are below the 0.05 threshold. Consequently, we reject the null hypothesis for all cases, providing compelling evidence that the observed differences are indeed statistically significant.

4. CONCLUSION & FUTURE WORK

This Enriched approach aims to increase the rating matrix based on users with different interests and preferences. To evaluate our methods, we compared it to the traditional approach using MovieLens dataset. Our approach outperforms traditional approaches and improves prediction accuracy. This confirms that the Enriched method yields better results than the traditional method. The achieved results of the current study show that innovation in the recommendation systems can be achieved by efficiency and combination of new and old approaches as the trend reveals in terms of research.

Our ongoing research concentrates on the exploration of online machine learning approaches, an emerging field that holds the potential to provide essential insights into the dynamics of evolving consumer bases. In our future work, we plan to delve into additional outlier detection techniques to further examine the evolution of user profiles over time.

REFERENCES:

[1] J. Ben Schafer, J. A. Konstan, and J. Riedl, “Ecommerce recommendation applications,” Data Min. Knowl. Discov., vol. 5, no. 1–2, pp. 115–153, 2001, doi: 10.1007/9781461516279_6.

[2] G. Linden, B. Smith, and J. York, “Amazon.com recommendations: Itemtoitem collaborative filtering,” IEEE Internet Comput., vol. 7, no. 1, pp. 76–80, 2017, doi: 10.1109/MIC.2003.1167344.

[3] C. A. Gomezuribe and N. Hunt, “The Netflix Recommender System_Algorithms, Business Value.pdf,” vol. 6, no. 4, 2015, doi: 10.1145/2843948.

[4] O. Celma, “Music recommendation and discovery in the long tail,” Citeulikeorg, p. 252, 2008, [Online]. Available: http://www.citeulike.org/user/bemike/article/4099800

[5] J. Callan et al., “Personalisation and Recommender Systems in Digital Libraries Joint NSFEU DELOS Working Group Report,” Library (Lond)., vol. 5, no. May, pp. 299–308, 2003, doi: 10.1007/s0079900401001.

[6] M. D. Ekstrand, “Collaborative Filtering Recommender Systems,” Found. Trends® Human–Computer Interact., vol. 4, no. 2, pp. 81–173, 2011, doi: 10.1561/1100000009.

[7] G. Jawaheer, M. Szomszor, and P. Kostkova, “Comparison of implicit and explicit feedback from an online music recommendation service,” pp. 47–51, 2006, doi: 10.1145/1869446.1869453.

[8] Y. Hu, Y. Koren, and C. Volinsky, “Collaborative Filtering for Implicit Feedback Datasets,” Gastroenterology, vol. 1), p. S415, 2008, doi: 10.1109/ICDM.2008.22.

[9] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, “Recommendation systems: Principles, methods and evaluation,” Egypt. Informatics J., vol. 16, no. 3, pp. 261–273, 2015, doi: 10.1016/j.eij.2015.06.005.

[10] A. Paterek, “Improving regularized singular value decomposition for collaborative filtering,” Proc. KDD cup Work., pp. 2–5, 2007, doi: 10.1145/1557019.1557072.

[11] S. Agrawal and J. Agrawal, “Survey on anomaly detection using data mining techniques,” Procedia Comput. Sci., vol. 60, no. 1, pp. 708–713, 2015, doi: 10.1016/j.procs.2015.08.220.

[12] Y. Zheng, M. Agnani, and M. Singh, “Identifying grey sheep users by the distribution of user similarities in collaborative filtering,” RIIT 2017 Proc. 6th Annu. Conf. Res. Inf. Technol., pp. 1–6, 2017, doi: 10.1145/3125649.3125651.

[13] P. Wei, Y. Li, Z. Zhang, T. Hu, Z. Li, and D. Liu, “An optimization method for intrusion detection classification model based on deep belief network,” IEEE Access, vol. 7, pp. 87593–87605, 2019, doi: 10.1109/ACCESS.2019.2925828.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[14] J. Breese, D. Heckerman, and C. Kadie, “Empirical analysis of predictive algorithms for collaborative filtering,” Proc. 14th Annu. Conf. Uncertain. Artif. Intell., pp. 4352, 1998, doi: 10.1111/j.15532712.2011.01172.x.

[15] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, “Recommender systems survey,” KnowledgeBased Syst., vol. 46, pp. 109–132, 2013, doi: 10.1016/j.knosys.2013.03.012.

[16] S. Kant, T. Mahara, V. Kumar Jain, D. Kumar Jain, and A. Kumar Sangaiah, “LeaderRank based kmeans clustering initialization method for collaborative filtering,” Comput. Electr. Eng., vol. 69, pp. 598–609, 2018, doi: 10.1016/j.compeleceng.2017.12.001.

[17] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin, “Combining contentbased and collaborative filters in an online newspaper,” Proc. ACM SIGIR ’99 Work. Recomm. Syst. Algorithms Eval., no. June, p., 1999.

[18] P. Adamopoulos, “On discovering nonobvious recommendations: Using unexpectedness and neighborhood selection methods in collaborative filtering systems,” WSDM 2014 Proc. 7th ACM Int. Conf. Web Search Data Min., pp. 655–659, 2014, doi: 10.1145/2556195.2556204.

[19] A. Srivastava, P. K. Bala, and B. Kumar, “New perspectives on gray sheep behavior in Ecommerce recommendations,” J. Retail. Consum. Serv., vol. 53, no. September 2018, p. 101764, 2020, doi: 10.1016/j.jretconser.2019.02.018.

[20] S. Ghorbani and A. H. Novin, “An Introduction on Separating GraySheep Users in Personalized Recommender Systems Using Clustering Solution,” Int. J. Comput. Sci. Softw. Eng., vol. 5, no. 2, pp. 14–18, 2016.

[21] M. A. Ghazanfar and A. Prugelbennett, “Fulfilling the Needs of GraySheep Users in Recommender Systems , A Clustering Solution”.

[22] A. Srivastava, “Gray sheep, influential users, user modeling and recommender system adoption by startups,” J Intell Inf. Syst, vol. 20, no. no.2, pp. 137–148, 2014, doi: 10.1145/2959100.2959103.

[23] S. J. Gong, H. Wu Ye, and H. S. Tan, “Combining memorybased and modelbased collaborative filtering in recommender system,” Proc. 2009 PacificAsia Conf. Circuits, Commun. Syst. PACCS 2009, pp. 690–693, 2009, doi: 10.1109/PACCS.2009.66.

[24] R. Prabhu, P. Shetty, D. R. Shwetha, and R. Hegde, “A review : Recommender System using Collaborative Filtering and Gray Sheep Problem,” vol. 6, no. 2, pp. 440–443, 2018.

[25] G. Guo, H. Qiu, Z. Tan, Y. Liu, J. Ma, and X. Wang, “Resolving data sparsity by multitype auxiliary implicit feedback for recommender systems,” KnowledgeBased Syst., vol. 138, pp. 202–207, 2017, doi: 10.1016/j.knosys.2017.10.005.

[26] S. Natarajan, S. Vairavasundaram, S. Natarajan, and A. H. Gandomi, “Resolving data sparsity and cold start problem in collaborative filtering recommender system using Linked Open Data,” Expert Syst. Appl., vol. 149, 2020, doi: 10.1016/j.eswa.2020.113248.

[27] H. Hwangbo and Y. Kim, “An empirical study on the effect of data sparsity and data overlap on cross domain collaborative filtering performance,” Expert Syst. Appl., vol. 89, pp. 254–265, 2017, doi: 10.1016/j.eswa.2017.07.041.

[28] N. Mishra, S. Chaturvedi, V. Mishra, R. Srivastava, and P. Bargah, “Solving sparsity problem in ratingbased movie recommendation system,” Adv. Intell. Syst. Comput., vol. 556, pp. 111–117, 2017, doi: 10.1007/9789811038747_11.

[29] S. Zhang and X. Wu, “Fundamentals of association rules in data mining and knowledge discovery,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 1, no. 2, pp. 97–116, 2011, doi: 10.1002/widm.10.

[30] J. Jooa, S. Bangb, and G. Parka, “Implementation of a Recommendation System Using Association Rules and Collaborative Filtering,” Procedia Comput. Sci., vol. 91, no. Itqm 2016, pp. 944–952, 2016, doi: 10.1016/j.procs.2016.07.115.

[31] M. G. Vozalis and K. G. Margaritis, “Using SVD and demographic data for the enhancement of generalized Collaborative Filtering,” Inf. Sci. (Ny)., vol. 177, no. 15, pp. 3017–3037, 2007, doi: 10.1016/j.ins.2007.02.036.

[32] M. Yu, T. Quan, Q. Peng, X. Yu, and L. Liu, “A modelbased collaborate filtering algorithm based on stacked AutoEncoder,” Neural Comput. Appl., vol. 34, no. 4, pp. 2503–2511, 2022, doi: 10.1007/s00521021059338.

[33] Simo Funk, “Netflix update: Try this at home,” https://sifter.org/simon/journal/20061211.html.

[34] G. Linden, B. Smith, and J. York, “Amazon.com recommendations: Itemtoitem collaborative filtering,” IEEE Internet Comput.,

vol. 7, no. 1, pp. 76–80, 2003, doi: 10.1109/MIC.2003.1167344.
[35] Brandyn Webb, "Netflix update: Try this at home."
[36] H. Chen et al., "A spatiotemporal estimation method for hourly rainfall based on FSVD in the recommender system," Environ. Model. Softw., vol. 144, p. 105148, 2021, doi: 10.1016/j.envsoft.2021.105148.
[37] Y. Zhu, X. Shen, and C. Ye, "Personalized Prediction and Sparsity Pursuit in Latent Factor Models," J. Am. Stat. Assoc., vol. 111, no. 513, pp. 241–252, 2016, doi: 10.1080/01621459.2014.999158.
[38] A. K. Ghosh and A. Chakraborty, "Use of EM algorithm for data reduction under sparsity assumption," Comput. Stat., vol. 32, no. 2, pp. 387–407, 2017, doi: 10.1007/s0018001606573.
[39] B. Kumar, A. Srivastava, and P. Kumar, "Cosine Based Latent Factor Model for Precision Oriented Recommendation," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 1, pp. 451–457, 2016, doi: 10.14569/ijacsa.2016.070161.
[40] X. Hu, Z. Mai, H. Zhang, Y. Xue, W. Zhou, and X. Chen, "A hybrid recommendation model based on weighted bipartite graph and collaborative filtering," Proc. 2016 IEEE/WIC/ACM Int. Conf. Web Intell. Work. WIW 2016, pp. 119–122, 2017, doi: 10.1109/WIW.2016.4.
[41] A. Gautam, P. Chaudhary, K. Sindhwani, and P. Bedi, "CBCARS: Content boosted contextaware recommendations using tensor factorization," 2016 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2016, pp. 75–81, 2016, doi: 10.1109/ICACCI.2016.7732028.
[42] A. K. Sahoo, C. Pradhan, and B. S. P. Mishra, "SVD based privacy preserving recommendation model using optimized hybrid itembased collaborative filtering," Proc. 2019 IEEE Int. Conf. Commun. Signal Process. ICCSP 2019, pp. 294–298, 2019, doi: 10.1109/ICCSP.2019.8697950.
[43] Q. Shambour, "A UserBased MultiCriteria Recommendation Approach for Personalized Recommendations," vol. 14, no. 12, pp. 657–663, 2016.
[44] R. Sharma, D. Gopalani, and Y. Meena, "Collaborative Filtering – Based Recommender System : Approaches and Research Challenges," pp. 1–6, 2017.
[45] A. El Fazziki, Y. E. M. El Alami, J. Elhassouni, O. El Aissaoui, and M. Benbrahim, "Employing opposite ratings users in a new approach to collaborative filtering," Indones. J. Electr. Eng. Comput. Sci., vol. 25, no. 1, pp. 450–459, 2022, doi: 10.11591/ijeecs.v25.i1.pp450459.
[46] A. El Fazziki, Y. El Madani El Alami, O. El Aissaoui, Y. El Allioui, and M. Benbrahim, Improving Collaborative Filtering Approach by Leveraging Opposite Users, vol. 1102 AISC, no. January. Springer International Publishing, 2020. doi: 10.1007/9783030366537_14.
[47] A. El Fazziki, O. El Aissaoui, Y. E. M. El Alami, Y. El Allioui, and M. Benbrahim, "A new collaborative approach to solve the graysheep users problem in recommender systems," 2019 3rd Int. Conf. Intell. Comput. Data Sci. ICDS 2019, pp. 1–4, 2019, doi: 10.1109/ICDS47004.2019.8942256.
[48] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the stateoftheart and possible extensions," IEEE Trans. Knowl. Data Eng., vol. 17, no. 6, pp. 734–749, 2005, doi: 10.1109/TKDE.2005.99.