



A Personalized Recommender System

Akshit Nassa , Shubham Gupta, Pranjal Jindalm, Achin Jain, P. Singh Lamba

Bharati Vidyapeeth's College of Engineering, INDIA

Emails: akshittnassa412@gmail.com; shubham.gupta1704@gmail.com; pranjaljindalpj@gmail.com ;
achin.mails@gmail.com; singhs.puneet@gmail.com

* Correspondence: achin.mails@gmail.com

Abstract

Due to social media, e-commerce, and the broader digitization of businesses, a data surge has occurred during the previous decade. The information is used to make informed decisions, forecast market trends, and identify patterns in consumer preferences. Following the widespread adoption of internet services, recommendation systems have become commonplace. The idea is to use filtering algorithms to recommend products to users who might be interested in them. Users are given recommendations for a media item such as movies by discovering user profiles of people who share similar interests. The preferences of users are first determined by allowing them to rate movies of their choosing. After some time, the recommender system will be able to better understand the user and recommend films that are more likely to get higher ratings. It also considers the impact of personal and situational factors on the user experience. In comparison to previous models, the experimental findings on the TMDB dataset provide a dependable model that is precise and generates more customized movie recommendations.

Keywords: Recommender system; Movie recommendation; filtering techniques; Dataset; Personalization; User Experience

1. Introduction

A recommendation system is a type of information filtering system that is tasked with assuming a user's preferences and making recommendations based on those priorities. The user has access to a wide range of recommendation system applications. Recommendation systems have grown in popularity over time and have lately been integrated into practically all online sites that consumers use. Films, podcasts, novels, and videos have different content than colleagues and stories on social media, commodities on e-commerce websites, and individuals on commercial and dating websites. Twitter, for example, can analyze your interactions with various stories on your wall to determine what types of stories appeal to you. These systems may often be improved by combining the efforts of a large number of people. Users have come to expect good outcomes as recommender systems have progressed. They have a disadvantage when it comes to services that are unable to provide appropriate recommendations. As a result, technical firms place a significant value on enhancing their recommendation structures[1]. The situation, however, is more intricate than it appears; each user has different preferences. Furthermore, even a single customer's taste might vary significantly based on a variety of factors, such as the user's mood, season, or type of activity[2].

1.1 Content-Based Recommender System

Content-based recommender systems require the end-user to describe using natural language or any other means like rating to show their preference[4-8]. This content obtained from the user's history uses to make a recommendation [9-11]. Each potential recommendation compares with the preferences of a user from the past and products that are most compatible with the perceived taste of the user recommends.

1.2 Collaborative filtering Recommender System

Because the noticed ratings are typically closely related across numerous individuals and things, the fundamental approach of collaborative filtering systems is that these undecided ratings can be credited. Take, for instance, two individuals named Ramu and Shamu who have very identical likes. If the ratings, which both have mentioned, are pretty close, then the fundamental algorithms can assess their similarity. In such instances, there is a good chance that the ratings, even if only one of them has a definite value, will be identical. This resemblance can utilize to create inferences about values that are only partially articulated. For the computation technique, almost all collaborative filtering programs emphasize leveraging either item associations or user associations—both types of correlation used in many models. In addition, some mock-ups use carefully developed optimization processes to produce a training model, similar to how a classifier generates a training model from the information mentioned or supplied.

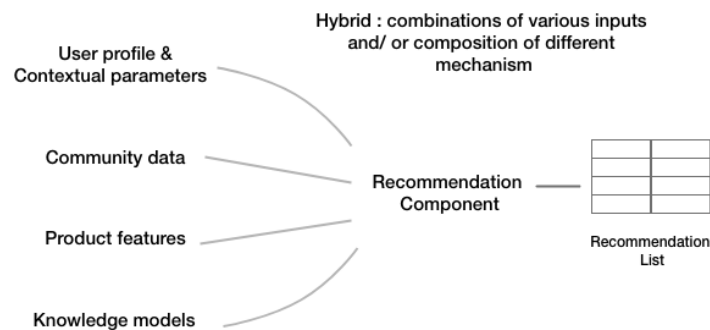


Figure 1: Different stellar recommendation

1.3 Hybrid Recommender System

To tailor the implementation, a hybrid recommender system combines two or more of the preceding strategies. They are typically used to improve the accuracy of the application's recommendations. Hybridization can achieve in several ways. The challenges traditional RSS face and their limitations are discussed below. The first challenge that is to tackle is finding neighbors. Given a large number of movies, a vast item profile, and a large number of users, the system must find recommendations extremely fast if predictions are to be made in real-time. There are two phases of computation required for the recommendation. In the first phase, the neighbors of the user to whom the recommendations are made are identified. In this phase, the similarity is computed between the targeted user and its neighbors. In the second phase, similarities between target users and the neighbors are computed and collected to make the prediction. After the prediction is computed, items that only cross a particular threshold will be recommended to the target user. The above two tasks involve complex processes, and hence, it is important to have efficient algorithms to complete the Tasks. The second challenge is to compute the rating accurately. Information about the neighbors is

needed before the predictions can be made. However, identifying the neighbors is a challenging task. User ratings are highly individualistic i.e., they can vary from user to user, and to identify similar users to form a cluster of neighbors is a challenging task. The third challenge is the issues of cold start and data sparsity. Recommended systems might not provide accurate predictions when there is not a significant amount of data to work with. Due to the scantiness of the available data, neighbor selection might become difficult. The fourth challenge traditional RSS must face is that the weight of niche movies is not factored into the prediction evaluation. Two movie watchers with only one niche movie in common have more similar tastes than those with only one mainstream movie in typical. In this paper, an improvement to the existing approach is presented, and various challenges are addressed.

2. Related work

Collaborative filtering recommender system is a topic in which much research is being conducted across the world. The first recommender systems were proposed by researchers in the late 1990s [15]. The idea of recommender systems emerged from the Tapestry project in 1992 [3]. Ever since it has been an active area of interest for data scientists because of the plethora of problems in the field and innumerable avenues of application in the real world. “Recommender systems root from extensive work in cognitive science, approximation theory, information retrieval, forecasting theories, and even market analysis and predictions” [15]. Content-based algorithms approach making suggestions as a query problem. They look at the history of ratings, descriptions of products by the user, and products the user have searched for. Table 1 summarises the related work.

Table 1: Related Work

Application Domain	Filtering Techniques	Related Research Papers
E-government	Knowledge-based	[17] “Meo, P.D., Quattrone, G. and Ursino, D. (2008)”
	Collaborative, Hybrid, and Knowledge-based	[18] “Lu, J., Shambour, Q., Xu, Y., Lin, Q. and Zhang, G. (2010)”
E-commerce	Knowledge-based, demographic	[19] “Garfinkel, R., Gopal, R., Tripathi, A. and Yin, F. (2006)”
	Knowledge-based; content based	[20] “Burke, R. (1999)”
E-learning	Content-based, collaborative, Hybrid and Knowledge-based	[8] “Balabanovic, M. and Shoham, Y. (1997)” [7] “Serrano-Guerrero, J., Herrera-Viedma, E., Olivas, J.A., Cerezo, A. and Romero, F.P.”

3. Information retrieval techniques in RSS

The digital world has been stuffed with unlimited data thanks to a variety of information sources. Because of the interactive engagement of people, the scenario has been amplified. To provide an efficient and successful recommendation, the RS must investigate all feasible dealing zones to collect and analyze relevant data to comprehend people's preferences and interests.

3.1 Cosine Similarity

Cosine similarity is a method to measure the difference between two non-zero vectors of an inner product space. The cosine similarity will measure the similarity between these two vectors, which is how identical the preferences between these two people are. Figure 2 shows The angle between $\mathbf{v1}$ and $\mathbf{v2}$, as seen in the cosine similarity diagram. The greater the similarity between the two vectors, the smaller the angle between them. It indicates that if the

angle between two vectors is small, they are almost identical, and if the angle is large, the vectors are substantially different from one another.

$$\text{Similarity} = \cos\theta = \frac{b.c}{\|b\| \|c\|} \tag{1}$$

Where,

$b.c \Rightarrow$ Is the Dot product of the two vectors

$\|b\| \|c\| \Rightarrow$ Is the product of each vector's magnitude

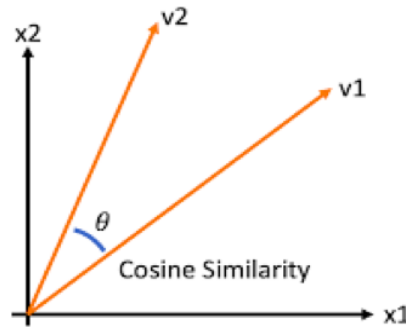


Figure 2: Cosine Similarity Diagram

3.2 TF-IDF

A fairly standard approach for this problem is to use a TF-IDF vectorizer. The TF-IDF algorithm uses to obtain essential words from the text in assessing the topic. The product of the term frequency, i.e., the number of times a given term (genre) appears in a document (movie genres), and the correct side term, which scales the term frequency based on the number of times a particular term appears in all documents, is what we have here (movies).

$$W_{i,j} = tf_{i,j} \cdot \log\left(\frac{N}{df_i}\right) \tag{2}$$

Where,

$tf_{i,j}$ =Number of Occurrences of i in j

$df_{i,j}$ =Number of Documents

N =Total Number of Documents

By assigning more significant weight to the less common genres, TF-IDF will assist capture the essential genres of each movie, something we would not get with, for instance, a Count Vectorizer. The next step is to obtain the TF-IDF weight. The TF and IDF vectors are combined to form a matrix. The weight of the TF-IDF is thus expressed as:

$$TF - IDF \text{ Weight} = TF(t, d) * IDF(t, D) \tag{3}$$

3.3 Weighted Rating

We will utilize the following criteria to create our Top Movies Chart and use a weighted rating method. Mathematically, the equation expresses as follows:

$$W = \frac{Rv + Cm}{v + m} \tag{4}$$

Where,

W = Weighted Rating

R = Average for the movie as a number from 0 to 10
 v = Number of Votes for the movies
 m = Minimum votes required to be listed in the Top 250
 C = Mean Vote across the whole report

The next step is choosing an acceptable value form, the number of votes needed in the chart. As a cutoff, we will pick the **60th percentile**. A film must receive more votes than at least 60% of the other films in the rankings.

3.4 Single Valued Decomposition (SVD)

The Singular Value Decomposition (SVD) is a linear algebra method widely utilized to solve problems in machine learning as a dimensionality reduction technique. When using the SVD technique, which is a matrix factorization method, a dataset's number of features is reduced from N to K ($K < N$). The SVD is a method used in collaborative filtering in recommender systems. It makes use of a matrix structure in the form of rows and columns, each column represents an item, and each row represents a user (movie). The matrix fills up with user-assigned values (ratings) for relevant items (movies)[16-20].

Matrix factorization ultimately tells us how aligned a user is with a collection of latent features and how well a movie fits into that set of latent features.

The singular value decomposition uses to factorize this matrix. It finds matrices' factors by factorizing a high-level matrix (user-item rating).


The SVD is a technique for reducing the number of variables in a problem by dividing a matrix into three smaller matrices as follows:

$$A = USV^T \quad (5)$$

As shown in the Figure down below: "Where A is a m x n utility matrix, U is a m x r orthogonal left singular matrix describing the relationship between users and latent factors, S is a r x r diagonal matrix describing the strength of each latent factor, and V is a r x n diagonal right singular matrix indicating the similarity between items and latent factors. The latent factors, in this case, are the characteristics of the items, for example, the genre of the movie" [9]. By removing the utility matrix A's latent factors, the SVD reduces its dimension. It creates an r-dimensional latent space for each user and item.

Let a vector \mathbf{x}_i represent each item, and a vector \mathbf{y}_u represent each user. Presumed rating by a user on an item \hat{r}_{ui} given as:

$$\hat{r}_{ui} = \mathbf{x}_i^T \cdot \mathbf{y}_u \quad (6)$$

 SVD is a type of factorization. \mathbf{y}_u can get the \mathbf{x}_i such that the square error between their dot product and the expected rating in the user-item matrix is as small as possible. It manifests itself in the following way:

$$\text{Min}(x, y) \sum_{(u,i) \in K} (r_{ui} - \mathbf{x}_i^T \cdot \mathbf{y}_u)^2 \quad (7)$$

A regularization element is included in the above formula as a penalty to allow the model to generalize well and not overfit the training data.

$$Min(x, y) \sum_{(u,i) \in K} (r_{ui} - x_i^T \cdot y_u)^2 + \lambda(|x_i|^2 + |y_u|^2) \tag{8}$$

A bias term is used in the algorithm to minimize the error between the value predicted by the model and the actual value. If for a user-item pair (u, i), μ is the average rating of all items, b_i is the average rating of item i minus μ , and b_u is the average user rating provided minus μ , the final equation may be written as:

$$Min(x, y, b_i, b_u) \sum_{(u,i) \in K} (r_{ui} - x_i^T \cdot y_u - \mu - b_i - b_u)^2 + \lambda(|x_i|^2 + |y_u|^2 + b_i^2 + b_u^2) \tag{9}$$

The above equation is the essential part of the algorithm for an SVD based recommendation system.

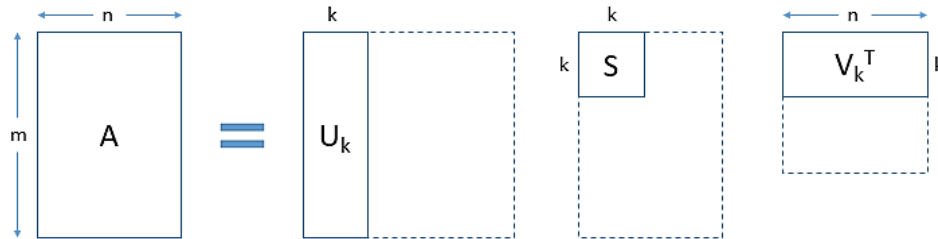


Figure 3: Matrix formation by dividing into small matrices

Assume r is the rank of $m \times n$ matrix A , so the SVD (Singular Value Decomposition) of matrix A is $A = USV^T$ in the above matrix formation. Singular Value Decomposition Thereinto, the columns of U are orthonormal. So are columns of V . S being a diagonal matrix, and elements on diagonal are the singular value of A .

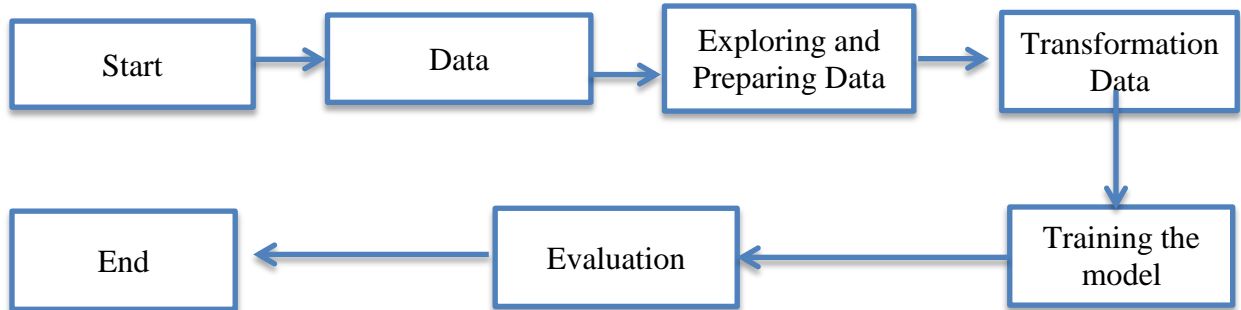


Figure 4: Flowchart of the algorithm used

4. Experiment Design and Discussion

4.1 Experimental Dataset

We obtained the original TheMovieDataBase from the Internet and put the dataset in a local database to test the algorithm in this work. The metadata for all 45,000 movies included in the Dataset is contained in these files. The movies in the dataset were released on or before July 1, 2017. The data points mention the cast, crew, storyline keywords, budget, revenue, posters, release dates, languages, production firms, countries, and TMDB vote counts and averages[12-15].

4.2 Experimental Evaluation Index

Prediction scores will differ depending on the similarity markers used. Two typical metrics for determining the accuracy of similarity approaches are mean absolute error (MAE), and root means square error (RMSE). “The more accurate the prediction, the lower the mean absolute error and root mean square error.” [14]. The below description shows the basic equation of MAE and RMSE.

- (1) **Mean absolute error (MAE)** is the average of the absolute error of the user has predicted score $r'_{u,i}$ and the actual score $r_{u,i}$ in the scoring test set:

$$MAE = \frac{1}{|p|} \sum_{(u,i) \in p} |r_{u,i} - r'_{u,i}| \quad (10)$$

- (2) **Root mean square error (RMSE)** is the mean square root of the actual score value $r_{u,i}$ and the predicted score value r' of the user in the test set:

$$RMSE = \sqrt{\frac{1}{|p|} \sum_{(u,i) \in p} (r_{u,i} - r'_{u,i})^2} \quad (11)$$

Where $|p|$ used in the above equation is equivalent to the size of the test set.

4.3 Experimental Program

The original TMDB dataset is divided into training sets and test sets to facilitate the evaluation of performance indicators. Further, the training set is used for model training, whereas the test set tests the pros and cons. Using cosine similarity, this paper tests the filtering algorithm based on the SVD, MAE, and RMSE indicators. This paper devised the following experimental approach based on the factors above:

Experiment: In the TMDB dataset, select a specified percent of training and test sets. The performance of the aforesaid method examines when the training set ratio alters to validate the influence on the recommendation effect using RMSE and MAE.

4.4 Experimental Analysis and result

We created a hybrid recommender system that combines the best of both worlds (content-based and collaborative filtering techniques) in this project. Table 2, and 3 shows the results.

- **Input:** The Title of a Movie and User ID.
- **Output:** Similar movies are sorted on the grounds of expected ratings by a particular user

Table 2: Result for hybrid (1, 'Se7en')

Title	Votes	Avg Vote	year	id	est	Title
Fight Club	9678	8.3	1999	550	4.81499	Fight Club
Zero Effect	56	6	1998	16148	3.164854	Zero Effect
Zodiac	2080	7.3	2007	1949	3.121718	Zodiac
The Game	1556	7.5	1997	2649	3.052226	The Game

The Girl with the Dragon Tattoo	2479	7.2	2011	65754	2.917327	The Girl with the Dragon Tattoo
Along Came a Spider	408	6.1	2001	2043	2.847169	Along Came a Spider
In the Valley of Elah	265	6.6	2007	6973	2.818515	In the Valley of Elah
The Curious Case of Benjamin Button	3398	7.3	2008	4922	2.797116	The Curious Case of Benjamin Button

Now the same is done with a different user.

Output: A different recommendation output is obtained for a different user i.e., personalized recommendation.

Table 3: Result hybrid (300, 'Se7en')

Title	Votes	Avg Vote	year	id	est
Fight Club	9678	8.3	1999	550	4.814993
The Game	1556	7.5	1997	2649	4.378067
The Social Network	3492	7.1	2010	37799	4.367137
The Girl with the Dragon Tattoo	2479	7.2	2011	65754	4.361135
Zodiac	2080	7.3	2007	1949	4.342165
The curious case of Benjamin Button	3398	7.3	2008	4922	4.223235
Zero Effect	56	6	1998	16148	4.125578

We get distinct recommendations for different users for the same movie using the above-mentioned hybrid recommender system. As a result, the recommendations are more bespoke and tailored to specific individuals, which improves overall user engagement and experience.

Experiment: To check the trend of decreasing RMSE and MAE we have trained our data on the test data which is some percentage of the total data set and apply the result of the training to the whole data set. The line chart drawn in Figures 7 and 8 corresponds to the table below (with different experimental values for different test set ratios). The specifics are provided in the figures below. When the training set ratio increases, the root mean square value of the Hybrid method suggested in this paper drops, resulting in a more tailored recommendation. As obtained in table 4, for the Ratio of training set = 10%, RMSE and MAE are equivalent to **0.9549** and **0.7453**, respectively. When the model is trained at a 100% training set ratio, RMSE and MAE came about to be **0.8651** and **0.6689**, decreasing the percentage error in RMSE and MAE, i.e., improving the accuracy of the algorithm by **9.4%** and **10.2%**

5. Conclusion

This research integrates social media engineering into hybrid recommendation technology and provides an algorithm based on SVD and TF-IDF. This research demonstrates a recommendation approach based on the above algorithms described in social media to tackle the challenges in traditional and suggested technologies. Experiments use to validate the suggested algorithm's performance. The experiment was designed utilizing The Movie Database dataset in this study. The SVD and the TF-IDF recommendation approach based on cosine similarity are evaluated using the performance indicators MAE and RMSE. The experimental results indicate that the propounded technique outperforms the matrix factorization-based recommendation in terms of performance.

Table 4: Experimental results for different training set ratios

The ratio of Training Set (%)	10	20	30	40	50	60	70	80	90	100
RMSE	0.9549	0.9345	0.9231	0.9193	0.9135	0.9101	0.9086	0.9063	0.8748	0.8651
MAE	0.7453	0.7236	0.7149	0.7099	0.7043	0.7022	0.7071	0.6946	0.6739	0.6689

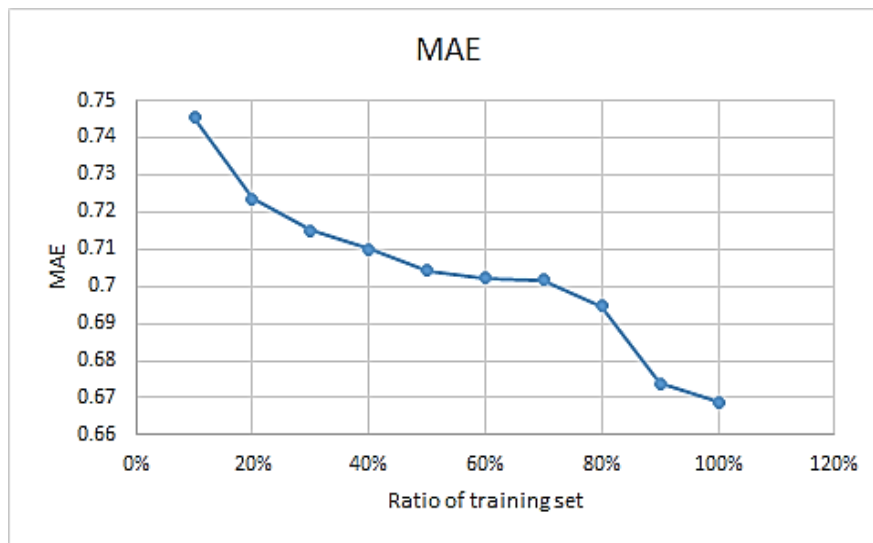


Figure 7: MAE changes for progressions in training set ratio

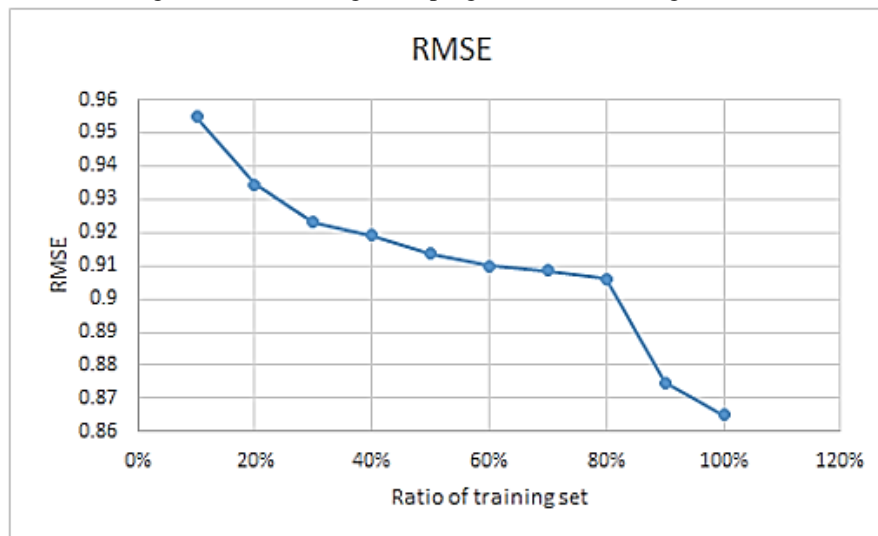


Figure 8: RMSE changes for different training set ratio

6. Future Work

As businesses attempt to make more use of their data (Moreover, as this field is relatively new, many improvements can be made over the existing model to suit the problem at hand more precisely and generate more accurate and personalized recommendations) more demographic data, along with nationality, race, location, mother tongue, and spoken languages, can help enhance the recommendations. The application can be made available on cross platforms to extend its reach and functionality. Instead of just recommending movies to the end-user, the application's functionality can be further expanded to finding and providing the OTT platforms of the respective movie, i.e., on what OTT platform the movie is available for streaming. Input from consolidated internet databases and authentic review sites like Rotten Tomatoes can be factored into the recommendation process.

References

- [1] Amatriain, X., Pujol, J. M., Tintarev, N., & Oliver, N. (2009, October). Rate it again: increasing recommendation accuracy by user re-rating. In *Proceedings of the third ACM conference on Recommender systems* (pp. 173-180).
- [2] Ansari A, Essegaier S, Kohli R. Internet Recommendation Systems. *Journal of Marketing Research*. 2000;37(3):363-375. doi:10.1509/jmkr.37.3.363.18779
- [3] Colombo-Mendoza, L. O., Valencia-García, R., Rodríguez-González, A., Alor-Hernández, G., & Samper-Zapater, J. J. (2015). RecomMetz: A context-aware knowledge-based mobile recommender system for movie show-times. *Expert Systems with Applications*, 42(3), 1202-1222
- [4] Guy, I., Zwerdling, N., Carmel, D., Ronen, I., Uziel, E., Yogev, S., & Ofek-Koifman, S. (2009, October). Personalized recommendation of social software items based on social relations. In *Proceedings of the third ACM conference on Recommender systems* (pp. 53-60)
- [5] Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000, December). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 241-250)
- [6] J. Stan, F. Muhlenbach, and C. Llargeron, "Recommender systems using social network analysis: challenges and future trends," in *Encyclopedia of Social Network Analysis and Mining*, pp. 1–22, Springer, New York, NY, USA, 2014
- [7] Serrano-Guerrero, J., Herrera-Viedma, E., Olivas, J. A., Cerezo, A., & Romero, F. P. (2011). A google wave-based fuzzy recommender system to disseminate information in University Digital Libraries 2.0. *Information Sciences*, 181(9), 1503-1516
- [8] Balabanović, M., & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3), 66-72
- [9] Christakou, C., Vrettos, S., & Stafylopatis, A. (2007). A hybrid movie recommender system based on neural networks. *International Journal on Artificial Intelligence Tools*, 16(05), 771-792
- [10] Santos, O. C., Boticario, J. G., & Pérez-Marín, D. (2014). Extending web-based educational systems with personalised support through User Centred Designed recommendations along the e-learning life cycle. *Science of Computer Programming*, 88, 92-109
- [11] Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). *Collaborative filtering recommender systems*. Now Publishers Inc
- [12] Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000, December). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 241-250)

- [13] Baudisch, P., & Terveen, L. (1999, May). Interacting with recommender systems. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems* (pp. 164-164)
- [14] Li, X., & Li, D. (2019). An improved collaborative filtering recommendation algorithm and recommendation strategy. *Mobile Information Systems, 2019*
- [15] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering, 17*(6), 734-749
- [16] Subramaniaswamy, V., Logesh, R., Chandrashekhar, M., Challa, A., & Vijayakumar, V. (2017). A personalised movie recommendation system based on collaborative filtering. *International Journal of High Performance Computing and Networking, 10*(1-2), 54-63
- [17] De Meo, P., Quattrone, G., & Ursino, D. (2008). A decision support system for designing new services tailored to citizen profiles in a complex and distributed e-government scenario. *Data & Knowledge Engineering, 67*(1), 161-184
- [18] Lu, J., Shambour, Q., Xu, Y., Lin, Q., & Zhang, G. (2010). BizSeeker: a hybrid semantic recommendation system for personalized government-to-business e-services. *Internet Research*.
- [19] Garfinkel, R., Gopal, R., Tripathi, A., & Yin, F. (2006). Design of a shopbot and recommender system for bundle purchases. *Decision Support Systems, 42*(3), 1974-1986
- [20] Burke, R. (1999, July). The wasabi personal shopper: A case-based recommender system. In *AAAI/IAAI* (pp. 844-849)