Enhancing Movie Recommendations: A Hybrid Filtering Approach Combining Collaborative and Content-Based Filtering

1st Esther Vemberly
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
esther.vemberly@binus.ac.id

2nd Eugenia Ancilla
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
eugenia.ancilla@binus.ac.id

3rd Anderies
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
anderies@binus.ac.id

4th Andry Chowanda
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
achowanda@binus.edu

2023 International Conference on Artificial Intelligence Robotics, Signal and Image Processing (AIRoSIP) | 979-8-3503-2747-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIRoSIP58759.2023.10873881

***Abstract*—The exponential growth of information and the widespread use of the internet has increased the importance of recommender algorithms, particularly in movie recommendation systems across various industries. By helping users select movies that match their interests, these systems significantly improve the movie-watching experience. However, a common challenge faced by these systems is the cold-start problem, which makes it difficult to provide accurate recommendations for new users with limited historical data. To address this issue and improve recommendation accuracy, this research paper presents a hybrid approach combining collaborative filtering (CF) and content-based filtering (CBF) to provide movie recommendations, where the model makes use of user ratings and movie datasets. The CF component identifies similar users and suggests popular movies, while the CBF component examines movie attributes such as titles, genres, ratings, and descriptions to recommend movies. The methodology involves data acquisition, cleaning, partitioning, and the design of collaborative, content-based, and hybrid models., where the dataset was partitioned into 60% for training and 40% for testing. The evaluation is based on the MovieLens 100K Dataset, using Root Mean Squared Error (RMSE) and Fraction of Concordant Pairs (FCP) metrics. Eventually, the result of the hybrid model obtains a training and testing RMSE of 0.881858 and 0.999145, respectively. For the FCP value, the model achieves a training value of 0.751256 and a testing value of 0.641928. Overall, the proposed hybrid approach highlights the importance of using CF and CBF approaches in recommender algorithms as it gives a balance between accuracy and diversity of the datasets and shows potential in addressing the cold-start problem and providing personalized recommendations. Future studies can focus on improving the hybrid model and exploring other evaluation metrics for a more thorough review.**

Keywords—Hybrid Filtering, Collaborative Filtering, Content-Based Filtering, Movie Recommendation System, Recommendation Algorithm

I. Introduction

Due to the exponential growth of information and the rapid use of the Internet, the importance of recommender algorithms in our daily lives is increasing [1]. A recommender algorithm is an artificial intelligence-based algorithm that examines user data and item characteristics in order to provide tailored suggestions or recommendations for items that match the user's preferences. Recommender systems are now widely and continuously used in various industries, including music, movies, e-commerce, and gadgets on demand. Movie recommendation systems have grown in significance with the immense number of movies currently accessible and the need to assist users in choosing movies that suit their interests [2]. Indirectly, these systems improve user engagement and activity while also helping to enhance the movie-watching experience.

Recommendation systems can provide personalized recommendations by analyzing data, utilizing algorithms, and possibly applying artificial intelligence (AI). To create an improved recommendation system, more information about a target user and like-minded people should be gathered and used. Over the years, various techniques have been proposed to carry out recommendations, including demographic filtering (DF), collaborative filtering (CF), and content-based filtering (CBF) [3]-[5]. However, a single approach may not be able to provide the most accurate recommendations. Many recommenders suffer from a cold-start problem in which it cannot invoke any reasoning for new users for whom they have not yet accumulated proper information. As a result, the need for a hybrid recommender that combines two or more filtering approaches in different ways has been proposed [6].

In this research paper, we proposed the implementation of a hybrid filtering approach, which combines CF and CBF to overcome the limitations of each approach and provide more accurate recommendations. Additionally, including side information on users and items alleviates the cold-start problem. The recommender utilizes user ratings and movie datasets to provide personalized recommendations to users. The CF part of the system analyzes the user ratings dataset to find similar users and recommends movies that are popular among those users. CBF relies on item attributes and characteristics to provide suggestions, so the CBF part analyzes movie datasets such as titles, genres, ratings, and descriptions to recommend movies similar in content to the user's previous preferences.

The outline of this paper is structured as follows. Section 2 discusses the related works in the implementation of hybrid recommendation systems for various kinds of datasets. Then, the methodology and our proposed model are given in Section 3. Followed by Section 4, which presents the result and the discussion of the proposed model. Finally, Section 5 presents the conclusion.

II. Related Works

Hybrid filtering is being applied in a growing number of recommendation systems. In this section, we review the existing literature on the topic of hybrid filtering, with a focus on the methodological techniques used in developing and evaluating various kinds of datasets. Related works have been implemented in datasets such as movies, images, smartphones, music, books, and many more.

TABLE I. E-commerce Item Datasets

Year	Method	Aim	Findings
2020	Content and user-item-based collaborative filtering [7].	To uniquely generate different types of weights, which help generate better and smaller recommendations.	For top-25 recommendations, the proposed approach significantly outperformed the usual user-based collaborative filtering method by 21% and the latest matrix factorization recommendation technique by 8%. By including users' context-related information in the recommendation system, the method may be further enhanced.
2023	A combination of a CF, CBF, and fuzzy expert system [8].	The aim of this article is to design a hybrid recommender system for proposing suitable items in an online store.	According to respondents, the proposed system by B. Walek and P. Fajmon may provide more suitable product suggestions according to the shopping cart compared to the online store Kosik.cz. The data presented indicates that the suggested recommender system produces positive results. The suggested system recommended the most items flagged as relevant when compared to other conventional recommender system methodologies.
2020	Using auxiliary information and a propoattention-based convolutional neural network [9].	- To put out a new hybrid PMF model that can distinguish between user preference and item attractiveness at the same time. - To conduct experiments on a variety of real-world datasets and demonstrate that it outperforms the baseline PMF, CLD, and modern algorithms.	Numerous real-world datasets, including three subsets of MovieLens and two subsets of Amazon, are used to assess extensive experiments. According to the encouraging results, the suggested DUPIA is better than the two baseline models and other cutting-edge methods regarding the RMSE criteria.

TABLE II. University-Related Datasets

Year	Method	Aim	Findings
2019	CF and CBF recommendation algorithm [10].	To make a prediction about the user's preference based on their interests, behaviors, and other data and to address the issue of difficult book choices and increase the rate at which library resources are used.	The proposed hybrid algorithm increased the recommendations algorithm's effectiveness and quality. Additionally, it may successfully address cold start. It helps prevent inefficient use of university book resources, increases the effectiveness of book recommendations, and improves the number of books borrowed.
2020	CF and CBF method [11].	To provide users with information that is personalized and tailored to their needs.	According to the experimental results, considering several criteria delivers improved outcomes, but since not all requirements are equally essential, it is important to investigate each one's relevance. Furthermore, using a hybrid system incorporating CF and CBF improves the results.

TABLE III. Cloud Services Datasets

Year	Method	Aim	Findings
2023	An improved user similarity measuring technique, clustering using K-medoid algorithms, and a weighted ranking aggregation model [12].	- To overcome the standard similarity calculation's poor accuracy and stability. - To increase the average satisfaction and quality of recommendations. - To conduct various experiments to confirm that the proposed technique is viable, especially in cloud manufacturing systems.	The HGRA technique outperforms the clustering-based algorithm and the ranking aggregation-based methods for group recommendation of both online and CMfg services under various group sizes. The findings demonstrated that the HGRA might, in addition to significantly enhancing the quality of group recommendations, also somewhat raise group satisfaction levels.
2020	Neighbor-based CF model improvement with factorization-based latent factor model integration [13].	- To extract the feature representations of users and IoT services, where co-occurrence embedding vectors for users and services are trained using the factorization method. - To oversee massive data and decrease security risk. - To perform external studies that verify the efficacy of the approach suggested by increasing recommendation effectiveness while ensuring user personal information.	SHCR exceeds the other three comparing approaches in terms of the Top-N recommendation regarding precision and recall. According to the findings, the suggested strategy improves prediction accuracy while also considering security concerns.

TABLE IV. OTHER DOMAIN DATASETS

Year	Method	Aim	Findings
2019	CBF, CF, and a complementarity-based recommendation method [14].	To provide an innovative hybrid system that will recommend Q&A documents to reduce overload.	The outcomes demonstrate that, in terms of the metrics, the suggested method performs better. The recommendation findings show that the experimental results with the knowledge requirements partitioning is better.
2020	A dynamic ensemble of CBF and CF methods depends on whether historical data, direct user preferences, or both exist [15].	To develop an assistant that utilizes historical data, if available, pre-processed data about the programs of radio stations, and user preferences regarding the type, content, tone, and scheduling of the programs. Additionally, the method can use user likes or ratings if they are available.	The proposed recommendation model obtains a precision of 0.5785. The approached results are satisfactory since they are consistent with those of relevant works that have been presented.

Table 1 - 4 demonstrates the application of hybrid recommendation systems in various datasets outside of movies, with each table showing different dataset domains according to the table name. These studies use a variety of methods, such as complementarity-based recommendation methods, collaborative filtering, and content-based filtering. The findings show improvements in recommendation performance, accuracy, and dealing with particular challenges within each dataset.

In this work, we proposed a recommendation system in the domain of movies, which are some of the frequently used datasets that use hybrid recommendation systems. This brings us to one of the hybrid recommender systems that use an expert system to suggest appropriate movies [16]. They suggest a hybrid recommender system that combines a fuzzy expert system to determine appropriate movies with CBF and CF method that uses the SVD algorithm. A movie hybrid recommendation system based on CBF and CF prediction using an artificial neural network is presented in one of the research projects to exceed the most recent techniques and increase the effectiveness of the conventional collaborative filtering approach [17]. Another researcher also makes use of the movie dataset and suggests a unique hybrid recommendation system based on multi-objective optimization to successfully balance accuracy, variety, originality, and coverage of recommendations at the same time. [18]. Similar to the one that we proposed, a hybrid recommendation system that combines content-based and collaborative filtering methods has been proven to produce an improved result in recommendation accuracy and root mean squared error (RMSE) [19]. It used a graph-based model that incorporates users' demographic and location information to improve the similarity of users' ratings.

Many approaches in hybrid recommendation systems have been proven to improve recommendations for many kinds of datasets and solve various kinds of problems regarding data sparsity, scalability, security, and others. In this work, we proposed a hybrid recommender in the domain of movies, consisting of a collaborative filtering part to consider users' preferences and content-based filtering to consider the movies' features to recommend personalized book recommendations. Our research objective also follows most of the literature reviews above, which is to solve the cold-start problems and give recommendations with high accuracy which outperforms the conventional filtering methods.

III. RESEARCH METHODOLOGY

The research methodology is divided into 7 phases: (1) Data Acquisition, (2) Data Cleaning and Preprocessing, (3) Partitioning the dataset into training and test data, (4) Design of the collaborative model, (5) Design of the content-based model, (6) Design of the hybrid model and (7) Evaluation of all the models (Figure 1).

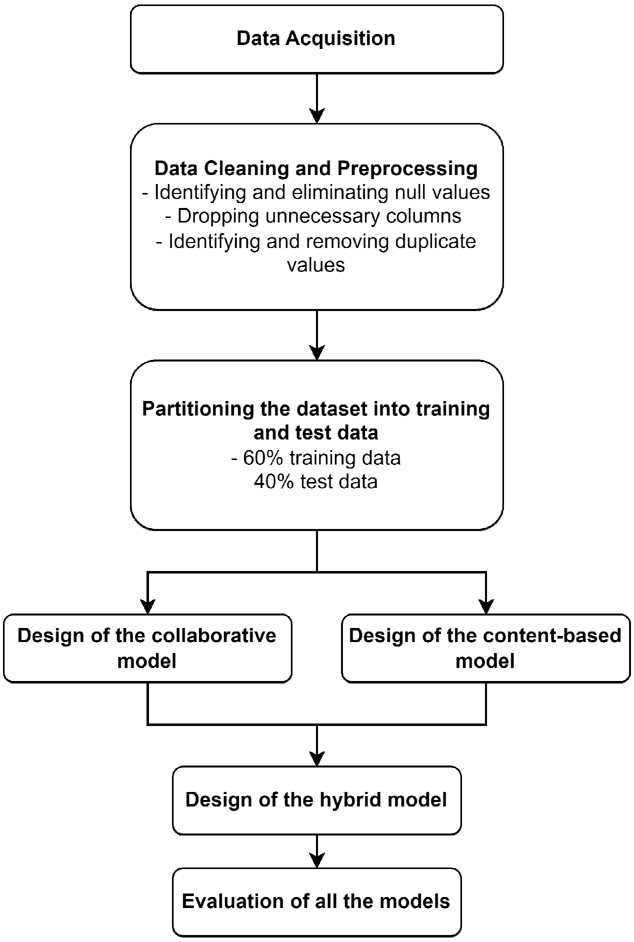
Data Acquisition
Data Cleaning and Preprocessing
- Identifying and eliminating null values
- Dropping unnecessary columns
- Identifying and removing duplicate values
Partitioning the dataset into training and test data
- 60% training data
40% test data
Design of the collaborative model
Design of the content-based model
Design of the hybrid model
Evaluation of all the models

Fig. 1. Research Methodology Phases.

A. Data Acquisition

We acquire and utilize the MovieLens 100K Dataset to build our recommendation system because it is freely available and frequently used to evaluate recommendation models.

B. *Data Cleaning and Preprocessing*

- Identifying and eliminating null values

  In the dataset, we verified how many values were null in each column and dropped them if there were any. However, there are no null values in the dataset.

- Dropping unnecessary columns

  In the table, we dropped the 'timestamp' column, as they are irrelevant to our research.

- Identifying and removing duplicate rows

  In the dataset, we verified if there were any duplicate values and dropped them if there were any. However, the dataset has no duplicate values, in which the code output shows 0 duplicate rows and 4 columns.

C. *Partitioning the MovieLens 100K dataset into training and test data*

To evaluate the performance of our recommendation models, we split our dataset into two subsets: 60% training data which is used to train our model, and 40% test data to test and evaluate our model.

D. *Design of the collaborative model*

The collaborative model analyses the behaviour and preferences of users to generate movie recommendations by taking into account the patterns and similarities in ratings among users.

E. *Design of the content-based model*

The content-based model analyzes the features and characteristics of movies, such as their genres, to recommend similar movies based on user preferences.

F. *Design of the hybrid model*

The hybrid model is designed by combining both collaborative and content-based approaches to provide more accurate and diverse recommendations.

G. *Evaluation of all the models*

In this final phase, all the designed models are evaluated using both training and test data. We used the Root Mean Squared Error (RMSE) and Fraction of Concordant Pairs (FCP) to assess the performance of each of our models. In the following subsections, we will present each of the two metrics in detail.

- Root Mean Squared Error (RMSE)

  In Eq. (1), the RMSE value is the average magnitude of the errors between the test dataset's actual ratings and the predicted ratings.

$$RMSE = \sqrt{\frac{1}{|\hat{R}|} \sum_{\widehat{r_{ui}} \in \hat{R}} (r_{ui} - \widehat{r_{ui}})^2} \tag{1}$$

  A lower RMSE indicates a higher accuracy.

- Fraction of Concordant Pairs (FCP)

  An issue for RMSE is that it does not take into consideration the rating scales that differ from one user to another. Thus, in addition, we used FCP to evaluate the algorithms using Eq. (2)-(4).

$$FCP = \frac{n_c}{n_c + n_d} \tag{2}$$

Where,

$$n_c = \sum |\{(i,j) \mid \widehat{r_{ui}} > \widehat{r_{uj}} \text{ and } r_{ui} > r_{uj}\}| \tag{3}$$

$$n_d = \sum |\{(i,j) \mid \widehat{r_{ui}} < \widehat{r_{uj}} \text{ and } r_{ui} < r_{uj}\}| \tag{4}$$

A higher value of FCP indicates a higher accuracy.

Those two-evaluation metrics help us determine which model or combination of models provides the best recommendation accuracy and effectiveness.

IV. Results and Discussions

This section presents the result of the evaluation of the three recommendation models that we designed: collaborative filtering, content-based filtering, and hybrid filtering. Root Mean Squared Error (RMSE) and Fraction of Concordant Pairs (FCP) are used to evaluate the performance of each model. Lower values in RMSE and higher values in FCP indicate better accuracy. We used bar graphs to visualize and provide a clear comparison of the results.

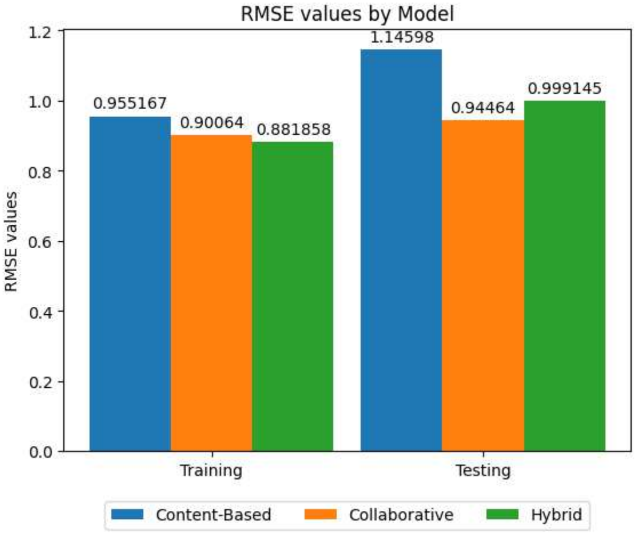

Fig. 2. RMSE Values of the CBF, CF, and Hybrid Model

A. *Collaborative Filtering Model*

The model achieved a training RMSE of 0.9006, indicating that the model fits the training data reasonably well and effectively captures the underlying patterns in the training dataset with a relatively low error level. However, the test RMSE of 0.9446 suggests a slightly higher error level in predicting movies for unseen data, which could indicate a slight overfitting of the model to the training data. The model also obtained an FCP value of 0.733433 in the training data, indicating a higher level of concordance with user preferences compared to the CB model. In the testing set, the value slightly decreases to 0.701847, which implies that the model maintains relatively consistent ranking accuracy.

B. Content-Based Filtering Model

The model achieved a training RMSE of 0.9552 which appears to have a reasonable fit to the training data. The test RMSE, on the other hand, suggests a higher error level, with an RMSE of 1.1460. This may be a result of limitations in the content-based filtering approach, such as its reliance on movie attributes and the lack of user-item interactions. The model also obtained an FCP value of 0.590981 which indicates that the model only captures moderate concordance with user preferences when ranking the movie dataset based on the movie's content similarities. In the testing set, the FCP value decreased to 0.440255, suggesting a lower ability to rank movies accurately for unseen user preferences.

C. Hybrid Filtering Model

Being the lowest among the three models, the training RMSE of 0.881858, as shown in Figure 2 suggests a good fit of the hybrid model to the training data, indicating a low degree of error. A little higher than the training RMSE, the test RMSE of 0.999145 demonstrates a respectably accurate prediction performance on the unobserved data. This is further emphasized by the FCP value of 0.751256 in the training set, which surpasses both CB and CF models, and the FCP value of 0.641928 in the testing set, which strikes a balance between those 2 models, as shown in Figure 3. The hybrid model achieves a trade-off between the 2 individual models, delivering moderate ranking accuracy and concordance in unseen scenarios.

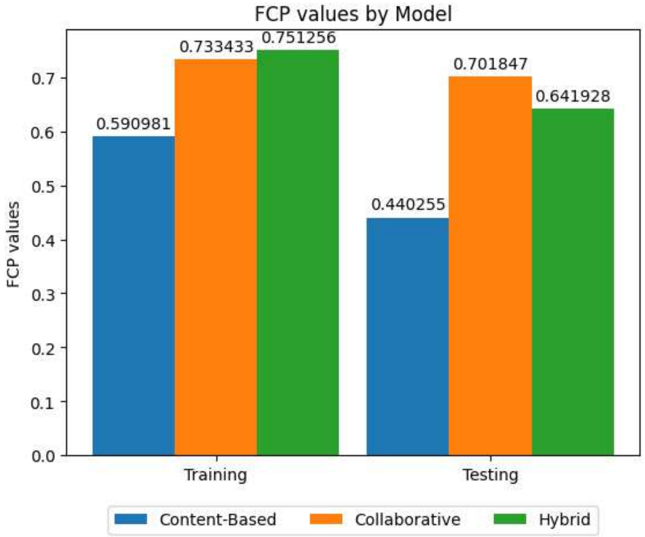

Fig. 3. FCP Values of the CBF, CF, and Hybrid Model

V. Conclusion

In this study, we investigated the performance of three recommendation models: collaborative filtering, content-based filtering, and hybrid filtering, for a movie recommendation system. The models were evaluated based on their training and test RMSE and FCP values. Although the test RMSE and FCP values of the hybrid model fall between the collaborative and content-based models, it nonetheless accurately predicted the movie recommendations for unseen data. The findings show that the hybrid filtering model offers a viable solution for movie recommendations. The model finds a balance between accuracy and diversity by applying the advantages of both collaborative and content-based filtering. The hybrid model showcased strong performance on the training data and delivered increased accuracy compared to the content-based model. Combining collaborative and content-based filtering shows that the hybrid model offers a more comprehensive approach to identifying user preferences and providing personalized recommendations.

Further research can focus on improving the hybrid filtering model, evaluating new features and algorithms, and examining the effects of other weighting schemes or fusion techniques. A more thorough evaluation of the recommendation models may be obtained by analysing additional metrics, including accuracy, recall, and coverage.

References

[1] J. A. Konstan and J. Riedl, "Recommender systems: From algorithms to user experience," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1–2. pp. 101–123, Apr. 2012. doi: 10.1007/s11257-011-9112-x.

[2] Z. Wang, X. Yu, N. Feng, and Z. Wang, "An improved collaborative movie recommendation system using computational intelligence," *J Vis Lang Comput*, vol. 25, no. 6, pp. 667–675, 2014, doi: 10.1016/j.jvlc.2014.09.011.

[3] A. Merve Acilar and A. Arslan, "A collaborative filtering method based on artificial immune network," *Expert Syst Appl*, vol. 36, no. 4, pp. 8324–8332, 2009, doi: https://doi.org/10.1016/j.eswa.2008.10.029.

[4] L.-S. Chen, F.-H. Hsu, M.-C. Chen, and Y.-C. Hsu, "Developing recommender systems with the consideration of product profitability for sellers," *Inf Sci (N Y)*, vol. 178, no. 4, pp. 1032–1048, 2008, doi: https://doi.org/10.1016/j.ins.2007.09.027.

[5] M. Jalali, N. Mustapha, Md. N. Sulaiman, and A. Mamat, "WebPUM: A Web-based recommendation system to predict user future movements," *Expert Syst Appl*, vol. 37, no. 9, pp. 6201–6212, 2010, doi: https://doi.org/10.1016/j.eswa.2010.02.105.

[6] M. Y. H. Al-Shamri and K. K. Bharadwaj, "Fuzzy-genetic approach to recommender systems based on a novel hybrid user model," *Expert Syst Appl*, vol. 35, no. 3, pp. 1386–1399, 2008, doi: https://doi.org/10.1016/j.eswa.2007.08.016.

[7] A. S. Tewari, "Generating Items Recommendations by Fusing Content and User-Item based Collaborative Filtering," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 1934–1940. doi: 10.1016/j.procs.2020.03.215.

[8] B. Walek and P. Fajmon, "A hybrid recommender system for an online store using a fuzzy expert system," *Expert Syst Appl*, vol. 212, Feb. 2023, doi: 10.1016/j.eswa.2022.118565.

[9] X. Zhang, H. Liu, X. Chen, J. Zhong, and D. Wang, "A novel hybrid deep recommendation system to differentiate user's preference and item's attractiveness," *Inf Sci (N Y)*, vol. 519, pp. 306–316, May 2020, doi: 10.1016/j.ins.2020.01.044.

[10] Y. Tian, B. Zheng, Y. Wang, Y. Zhang, and Q. Wu, "College library personalized recommendation system based on hybrid recommendation algorithm," in *Procedia CIRP*, Elsevier B.V., 2019, pp. 490–494. doi: 10.1016/j.procir.2019.04.126.

[11] A. Esteban, A. Zafra, and C. Romero, "Helping university students to choose elective courses by using a hybrid multi-criteria recommendation system with genetic optimization ☆," vol. 194, p. 105385, 2020, doi: 10.1016/j.knosys.

[12] J. Liu, Y. Chen, Q. Liu, and B. Tekinerdogan, "A similarity-enhanced hybrid group recommendation approach in cloud manufacturing systems," *Comput Ind Eng*, vol. 178, p. 109128, Apr. 2023, doi: 10.1016/j.cie.2023.109128.

[13] S. Meng, Z. Gao, Q. Li, H. Wang, H. N. Dai, and L. Qi, "Security-Driven hybrid collaborative recommendation method for cloud-based iot services," *Comput Secur*, vol. 97, Oct. 2020, doi: 10.1016/j.cose.2020.101950.

[14] M. Li, Y. Li, W. Lou, and L. Chen, “A hybrid recommendation system for Q&A documents,” *Expert Syst Appl*, vol. 144, Apr. 2020, doi: 10.1016/j.eswa.2019.113088.

[15] A. J. Fernández-García, R. Rodriguez-Echeverria, J. C. Preciado, J. Perianez, and J. D. Gutiérrez, “A hybrid multidimensional Recommender System for radio programs,” *Expert Syst Appl*, vol. 198, Jul. 2022, doi: 10.1016/j.eswa.2022.116706.

[16] B. Walek and V. Fojtik, “A hybrid recommender system for recommending relevant movies using an expert system,” *Expert Syst Appl*, vol. 158, Nov. 2020, doi: 10.1016/j.eswa.2020.113452.

[17] Y. Afoudi, M. Lazaar, and M. Al Achhab, “Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network,” *Simul Model Pract Theory*, vol. 113, Dec. 2021, doi: 10.1016/j.simpat.2021.102375.

[18] X. Cai, Z. Hu, P. Zhao, W. S. Zhang, and J. Chen, “A hybrid recommendation system with many-objective evolutionary algorithm,” *Expert Syst Appl*, vol. 159, Nov. 2020, doi: 10.1016/j.eswa.2020.113648.

[19] Z. Z. Darban and M. H. Valipour, “GHRS: Graph-based Hybrid Recommendation System with Application to Movie Recommendation,” Nov. 2021, doi: 10.1016/j.eswa.2022.116850.