The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-4/W3-2022
The 7th International Conference on Smart City Applications, 19–21 October 2022, Castelo Branco, Portugal

INVESTIGATING DIFFERENT SIMILARITY METRICS USED IN VARIOUS RECOMMENDER SYSTEMS TYPES: SCENARIO CASES.

Oumaima Stitini*, Soulaimane Kaloun, Omar Bencharef

*Corresponding author: Oumaima Stitini, oumaima.stitini@ced.uca.ma
Computer and System Engineering Laboratory, Cadi Ayyad University, Marrakesh, 40000, Morocco

KEY WORDS: Recommendation System RS, Similarity, Correlation, Distance, Item-based filtering, Collaborative Filtering, Content-based Filtering, Hybrid Filtering.

ABSTRACT:

A recommendation system represents a very efficient way to propose solutions adapted to customers needs. It allows users to discover interesting items from a large amount of data according to their preferences. To do this, it uses a similarity metric, which determines how similar two users or products are. In the case of recommender systems, similarity computation is a practical step. The calculation of similarity may be used for both items and users. Following the similarity calculation, a user or item with a comparable computation value can be recommended together with the goods to a user with similar preferences. The user's requirements influence the choice of similarity metric. This paper explores various similarity measurement methods employed in recommender systems. We compare correlation and distance techniques to determine the capabilities of different similitude calculation algorithms and synthesize which similarity measure is adapted for which type of recommendation.

1. INTRODUCTION

The advancement of Cloud computing, especially the World Wide Web, has been remarkable. There are new services on the computer as a result of this upgrade, such as papers, information, or stories to read, films to watch, or goods to buy. Making the best pick from these platforms vast products has become more challenging as e-commerce site development and Internet of Things use have grown. Users automatically gain from recommender system characteristics. Users rely on a film's storyline choice, whether it is expressed as communication data (genre, actors, or script) or satisfaction from watching the film's trailer. The emotional affinity of consumers with the film is significantly influenced by the media content. The use of an online recommendation system has grown commonplace (Stitini et al., 2021). Recommendation systems are a very important technological innovation that helps consumers identify items they want to order. Currently, most users make online purchases with a single click because it is easier and faster, and banking is also faster when done online. A recommendation system is a tool that helps end customers find the best items for their needs. For filtering category recommendations, these systems use statistical approaches and knowledge discovery techniques. There are mainly several approaches to generate recommendation of these systems [(Stitini et al., 2020)]: collaborative filtering, content-based filtering (Stitini et al., 2022), and hybrid filtering. Collaborative filtering is a popular and widespread approach to providing recommendations based on the expectation that users with similar preferences will have similar preferences in the future (Gazdar and Hidri, 2020). There are two primary approaches to recommend items in the collaborative filtering category: model-based recommendation and neighborhood-based recommendation called also memory-based filtering. Collaborative filtering, in general, uses a similarity scale to locate active user neighbors and standard components of a candidate (Fkih, 2021). The collaborative filtering algorithm method begins by collecting user information to construct a user profile or sample of forecasting jobs, including user attributes, behavior, or resource content (Suganeshwari and Ibrahim, 2018). The computation of similarities between items and users is the next stage. In the content-based filtering technique, attribute fields of people and products are gathered, and the most significant similarity score is taken into account. The primary focus of this research study is the examination of numerous similarity measurements and the standard features of similarity metrics, and which one to employ at which moment. In this research paper, further experimentations of the performance of the most commonly used similarity measures were also carried out. We conduct a precise comparative study between the two using the same dataset and evaluation measures. The further part of this study is composed as follows: The underlying theoretical knowledge is described in Section 2, and the prior research is reviewed in Section 3. Section **??** contrasts the assessed commonalities and examines the implementation environment, and section 5 concludes with a conclusion and suggests a future study.

2. PRELIMINARY KNOWLEDGE

The degree to which two items are alike is measured in numerical terms by their similarity. As a result, for pairings of items that are more similar, similarities are higher. In recommender system applications, selecting the appropriate similarity metric is critical. The representation of the objects, which might be in probabilistic or vector form and numeric or binary form, influences the choice of the similarity measure. Generally we can divide similarity metrics into two different groups as mentioned in the Figure 1.

2.1 Correlation Similarity Measurement

The linear link between variables is measured via correlation. It is a numerical measure of how similar two data items are. Items are closely similar if the number obtained is high.

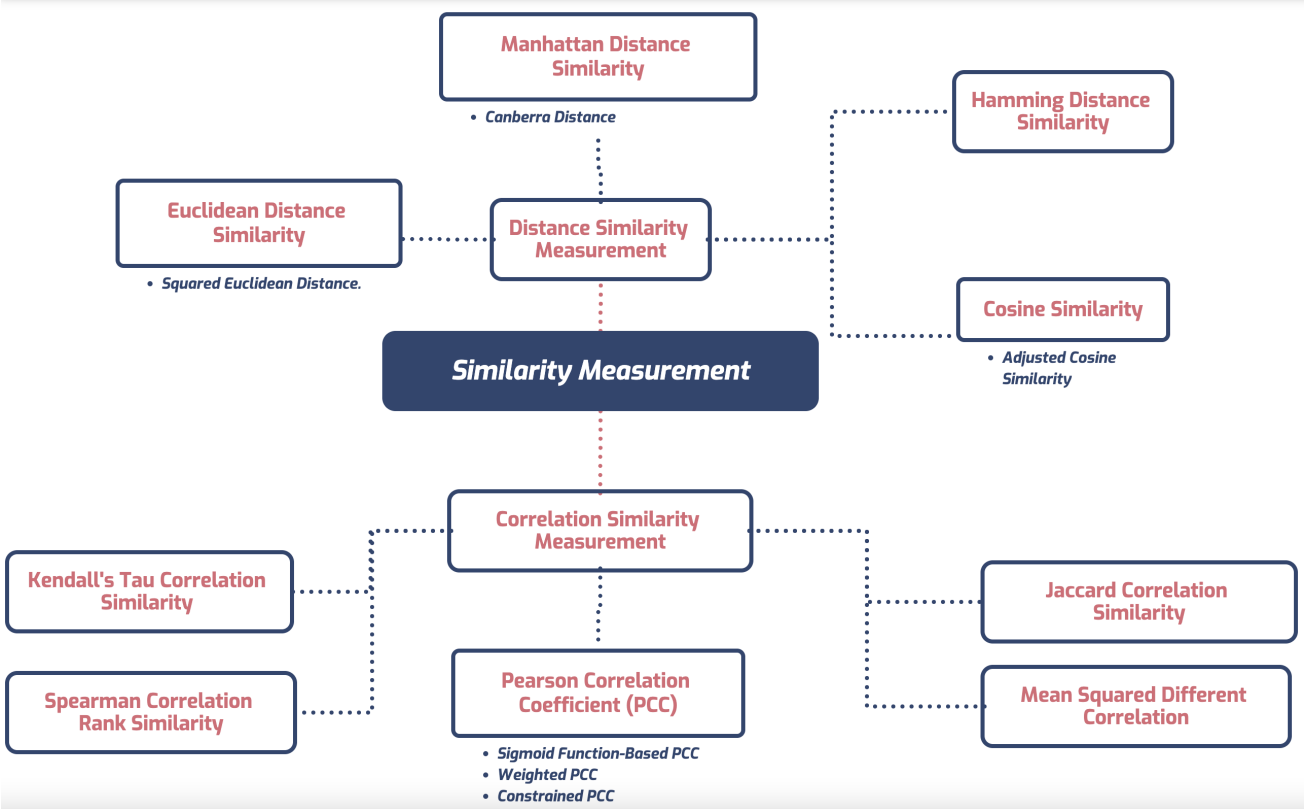

Figure 1. Similarity metrics types.

2.1.1 Pearson Correlation Coefficient PCC: Pearson's correlation coefficient (PCC) represents the linear correlation between users/objects. It is represented by the ratio of the covariation of two users and their standard deviation when only the co-rated items are involved.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

2.1.2 Spearman Rank Correlation Similarity In Spearman Rank Correlation, the similarity is calculated using rankings rather than ratings, and this approach eliminates the problem of normalizing rating. It is ineffective for incomplete orderings. Even if the ratings are comparable, there are many similarities.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (2)$$

2.1.3 Jaccard Correlation Similarity The Jaccard similarity coefficient, also known as the Tanimato coefficient similarity measure, is a prominent approach for measuring the similarity of users/items. Only the amount of shared ratings between the two users is taken into account when calculating similarity. Even if the ratings are comparable, there are many similarities.

$$JCS(u,v) = \frac{|I_u| \cap |I_v|}{|I_u| \cup |I_v|} \quad (3)$$

The Jaccard similarity coefficient is used to compare members of two groups to determine which are common and which are distinct.

2.1.4 Mean Squared Difference Correlation Absolute ratings are taken into account via Mean Squared Difference rather than the total amount of standard ratings. Then, similarity would be determined by averaging these squared differences; the smaller the mean squared difference, the more similar the two are.

2.1.5 Kendall's Tau Correlation Similarity Kendall's tau is a correlation coefficient that is quite close to Spearman's. Both of these measurements of a connection are non-parametric. The coefficients of Spearman and Kendall are derived using ranking data rather than raw data. Like Pearson's and Spearman's correlations, Kendall's Tau is always between -1 and +1, with -1 indicating a complete and negative link between two variables and 1 indicating a strong, positive relationship. To calculate Kendall's rank correlation coefficient (also known as Kendall's tau) instead of transforming the information to scores and afterwards estimating the Pearson correlation, $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. For any pair of indices: $1 \leq i < j \leq n$,

- We talk about concordant pair if both $x_i \geq x_j$ and $y_i \geq y_j$, or both $x_i \leq x_j$ and $y_i \leq y_j$. (i,j).
- We talk about discordant pair if both $x_i \geq x_j$ and $y_i \leq y_j$, or both $x_i \leq x_j$ and $y_i \geq y_j$. (i,j).

$$T_{x,y} = \frac{|concordantpairs| - |discordantpairs|}{n(n-1)/2} \quad (4)$$

2.2 Distance Similarity Measurement

2.2.1 Cosine Similarity The cosine similarity measure determines how semantically similar user-provided rating vectors are by computing the cosine angles that were formed between them. More similarity is implied by angles with lower values, and vice versa. The uncentered cosine similarity measure is so named since it does not provide for data centering or modification of preference values [(Zubair et al., 2019)]. Cosine similarity is computed as follows:

$$\cos(\mathbf{t},\mathbf{e}) = \frac{\mathbf{te}}{\|\mathbf{t}\|\|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2}\sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \quad (5)$$

2.2.2 Euclidean distance Similarity The Euclidean Distance Metric collects the most effective and often used distance measurements.

$$EDS\,(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (6)$$

2.2.3 Manhattan Distance Similarity The sum of absolute differences of Cartesian coordinates is used to calculate the distance between Manhattan and other locations.

$$MDS(p,q) = \sum_{i=1}^{n} |p_i - q_i| \quad (7)$$

2.2.4 Hamming Distance Similarity When two binary strings of equal length are matched, the Hamming distance is calculated. It is the number of bit locations where the two bits are not the same. It is mainly used for error detection and correction during data transmission across computer networks. In coding theory, it is also used to measure similarity.

$$HDS(i,j) = \sum_{i=1}^{n-1} [y_{i,k} \# y_{j,k}] \quad (8)$$

3. RELATED WORK

The use and availability of the recommendation system date back to the 1990s, and it has since become an indispensable tool for online shoppers and browsers. A recommendation system guides the user to discover the correct information or product. Collaborative filtering is the best way to develop automatic predictions about a user's preferences by analyzing preferred information from nearby users. Numerous similarity measurements can be used to locate the nearest user.

(Khojamli and Razmara, 2021) covers the many similarity metrics used in neighborhood-based collaborative filtering recommender systems and draws attention to other elements that affect how effective the suggestions are overall. They have suggested that organizing objects into categories would allow them to forecast ratings for all unrated items, increasing the density of the user-item matrix.

The authors of (al., 2019) have shown the explicit rating significance rather than just calculating distances among users using similarity measures. Many similarity measures are conducted, it is concluded that it is impossible to relate between users effectively since it provides relatively equivalent similarity values. They mention that better results are obtained from a combination of similarity measures because another measure strengthens the weakness of each of the measures. Authors in (Zubair et al., 2019) analyzed the disadvantages of the existing similarity measure. They compare all correlation similarities with the best distance similarity, and they obtained that the cosine similarity models performed better than other similar models.

4. RESEARCH METHODOLOGY

4.1 Scenario Case 1: Collaborative Recommender Systems

A well-liked recommendation algorithm called collaborative filtering bases its recommendations on the actions or evaluations of other system users as mentioned in Figure 3. The fundamental principle underlying this method is that information from other users may be chosen and compiled to offer reliable forecast regarding the specific user preferences. If the consumer concurs with the things quality and relevancy, they are likely to concur with other products. The suggestion list in user-based CF is produced by comparing the target user's choices with all other users with those interests.

Figure 2. Flowchart of the Collaborative Filtering approach.

4.1.1 User rating matrix Because observed ratings are frequently strongly linked across different users and objects, the fundamental principle behind collaborative filtering approaches is that it is possible to infer the ratings that a user has not explicitly given to the items. They rely on user opinions rather than any specific item attribute to function.

The matrix used to represent product ratings is called the user rating matrix URM. It serves as the primary input to a collaborative filtering system. Each column, i, denotes an object, and each row, u, denotes a user. The matrix's components, r, u, and i, explain the previous interaction between user u and object i as mentioned in Table 2.

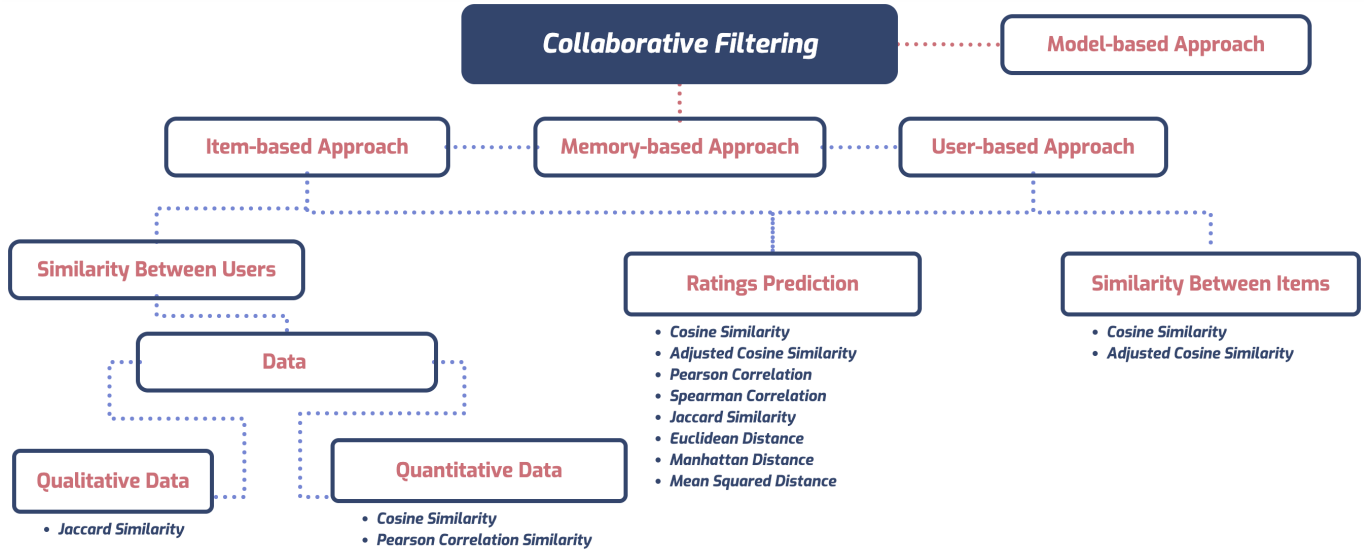

Figure 3. Similarity Metrics for Scenario Case 1.

Table 1. Most important metrics used in collaborative filtering.

Similarity Types	Why We Use	Why We Not Use	When We Use
Cosine Similarity	Good results: 0 indicates the dissimilarity otherwise the 1 indicates the high similarity.	Its preferable to not use this type when we want to calculate the dissimilarity.	When we wish to determine how similar a list of users or products is to one another.
Euclidean distance Similarity	Good results: 0 indicates a high similarity, otherwise all results up to 0 indicates a dissimilarity.	We use only in the case of dissimilarity.	When we want to calculate a dissimilarity between a list of users or items.
Manhattan Distance Similarity	-	-	taxi-distance
Hamming Distance Similarity	The number of locations at which the corresponding characters are different between two strings of equal length is known as the Hamming Distance. We can determine the Hamming distance since the lengths of these strings are equivalent.	Not use this type when we don't have the strings.	When we want to measure the similarity between two strings of the same length.

Table 2. User rating matrix representation.

	i_1	i_2	i_n
u_1	r_{11}	**?**	r_{1n}
u_n	**?**	r_{n2}	r_{nn}

4.1.2 Case Study: Movies Example Our example show a demonstration of 3 users who rated 4 movies as mentioned in Table 3. The sign ? means that the user has not rated yet the specific movie, other values are comprised between 1 and 5.

	i_1	i_2	i_3	i_3	i_4
u_1	4	2	?	5	4
u_2	5	3	4	?	3
u_3	3	?	4	4	3

Table 3. User movie rating example.

4.1.3 Similarity between users: Figure 2 represents all similarity measure that we can use inside a collaborative filtering recommender systems. Table 4 similarity between users, the highest is the similarity, the higher similar users are. Using Cosine Similarity mentions that both users are similar. Otherwise Hamming Distance Similarity indicated that both users are dissimilar. That's why we should always use Cosine Similarity when we will calculate similarity between users. Hamming Distance is used only when we want to compare between strings and not appropriate for user similarity.

Similarity types	Similarity values	
	u_2	u_3
Cosine Similarity	0,97	0.94
Euclidean distance Similarity	2.0	2.65
Manhattan Distance Similarity	4	5
Hamming Distance Similarity	0.25	0.5

Table 4. Calculated similarity between users.

4.1.4 Ratings Prediction: When we want to examine the rating of something which means linear correlation we should use one of correlation similarity measurements (Nudrat et al., 2022) already cited in the previous section. Otherwise in the case of similarity between users or between items we can use one of distance similarity measurement. The table 1 illustrates all criteria that we need before choosing the suitable distance similarity measure.

4.2 User-based collaborative filtering

The main principle of user-based collaborative filtering is to identify users who have similar tastes and to recommend the products those users value the most to them. Coming up with

a method to determine user similarities is the first challenge. We may compare user preferences based on ratings from users. When two people rate a variety of goods similarly, we may assume that they have similar opinions on those two things. Similarly, we may assume that two users are comparable if they share similar opinions on a wide range of diverse topics. Algorithm 1 illustrate steps used in the user-to-user algorithm.

Algorithm 1 User-based collaborative filtering algorithm.

Input : User preferences
Output : Recommendation
Calculate user similarity.
Locate users who are similar to users u.
Predict user ratings for items with no ratings
Recommendation.

4.3 Item-based collaborative filtering

For the purpose of making rating predictions, we must determine the set of items that are much more comparable to the target item using item-based techniques (Singh et al., 2020). According to how many individuals have assessed each pair of items, the objective is to ascertain how similar they are (Atashkar and Safi-Esfahani, 2020). The ratings supplied by the user for that item are then used to evaluate if the user would love the target items. Algorithm 2 illustrate steps used in the item-to-item algorithm.

Algorithm 2 Item-based collaborative filtering algorithm.

Input : User preferences
Output : Recommendation
Find similar things for item i.
Using the ratings for similar things, estimate the item I rating.
Can employ the same similarity measures and forecasting techniques as the user-user model.

4.4 Model-based collaborative filtering

In the model-based approach, users and things are represented by a utility matrix, which is then divided into an A and B matrix, where A stands for the user and B for the goods. Various methods are employed for the breakdown of the matrix. Each item receives a score, and the highest-scoring product is suggested. When there is a lot of data accessible, this model is beneficial. Three subtypes of this strategy are further separated:

- Technique utilizing clustering.
- Matrix factorization methods.
- Neural networks and deep learning.

Figures 8, 9, and 10 illustrate results for RMSE, MAE, and ratings for model-based collaborative filtering.

4.5 Scenario Case 2: Content-based Recommender Systems

By assessing the similarity of characteristics across objects, the system learns to generate suggestions. Based on the user's prior evaluations, a content-based recommendation algorithm provides a personal prototype. This prototype takes into account user choices and is flexible enough to address new issues. Content-based filtering is most commonly used for textual data, such as published articles, videos, or documents containing metadata. It is based on the principle that recommendation results depend on what the user has already viewed. Content-based filtering systems examine two things: items and user preferences to elaborate a model based on this information. They utilize a user's specific interests and try to match the properties of the numerous content items recommended with the user's profile.

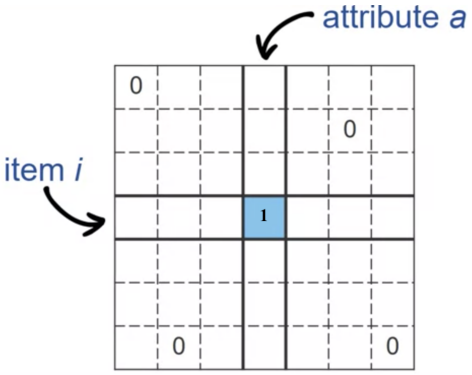

Figure 4. Item content matrix.

4.5.1 Item content matrix ICM Content-based filtering compare items based on their attributes. If a user expressed a preference for an item is likely to like similar items. If the item has that attribute, there is 1, otherwise 0 as mentioned in Figure 4.

4.5.2 Similarity metrics used in Content-based Filtering

Dot Product Dot Product similarity metric is the number of common attributes between two items. Figure 5 illustrates a scenario case for dot product.

$$S_{ij} = \vec{i} \times \vec{j} = \sum_{i=1}^{n} ij \tag{9}$$

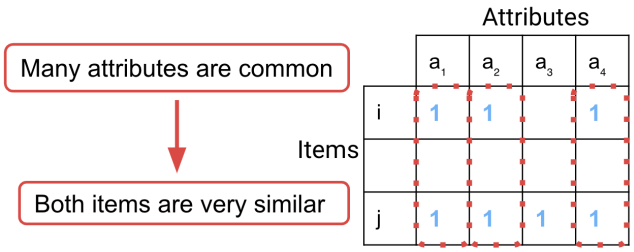

Figure 5. Scenario case for dot product similarity metrics.

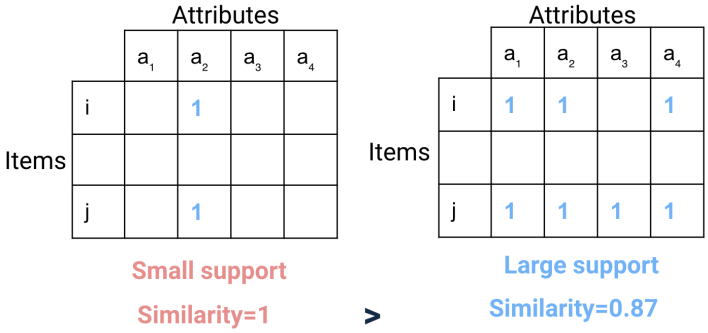

Figure 6. Scenario case for cosinus similarity metrics.

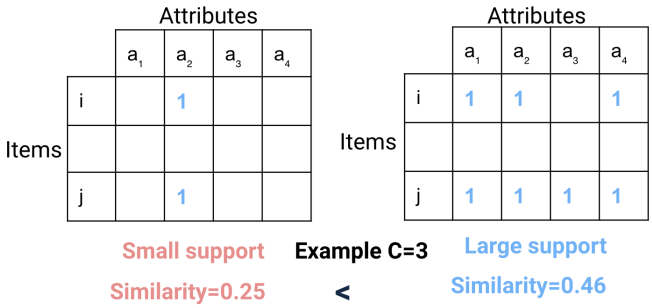

Figure 7. Scenario case for shrinking the cosinus similarity metrics.

Cosine similarity Cosine similarity is the normalization of the dot product. Figure 6 show an example for small and large support, the small support achieve the high similarity metrics even if we have just one common attribute, however the large support have low similarity in comparison with the small support. For that we should add the shrink term.

$$S_{ij} = \frac{\vec{i} \times \vec{j}}{|\vec{i}| \times |\vec{j}| + C} \tag{10}$$

Generally, shrinking reduce similarity to take into account only most similar with large support as mentioned in Figure 7.

4.6 Scenario Case 3: Hybrid Recommender Systems

Hybrid recommender systems, which integrate the content-based and collaborative filtering strategies to offer better reasonable estimations and get beyond each method's drawbacks. User tastes are ever-changing. A single content-based or collaborative filtering cannot give users highly accurate product recommendations. As a result, the suggestion of a product to the user in the Hybrid Recommendation System is based on a mix of content-based filtering and collaborative filtering. The similarity metrics that we should use in this last scenario case is the combination of both previous scenario cases.

5. CONCLUSION

In summary, many advances have been made in recommender systems to facilitate each user's wish. At the moment, the Artificial Intelligence algorithm has become the standard for designing systems that collect user preferences and other criteria and recommend to a specific person. Detecting similarity between users remains a primary task in a recommendation system. These systems are mainly used and implemented by large online applications. Concrete examples include friend suggestions on social applications such as Facebook, Twitter and LinkedIn, profile suggestions on Instagram, product suggestions on Amazon, and video recommendations. In e-commerce, recommender systems are a valuable tool for assisting customers in making purchases. The multiple similarity metrics employed in a neighbourhood-based collaborative filtering recommender system are discussed in this study. We examine similarity metrics to analyze the performance of different similarity calculation techniques and find the most appropriate measure. The main idea of this research paper is to categorize perfectlly each recommender system type with specific similarities metris that will be used. Regardless of the application area, this article gives a framework for finding related users or products. The originality of this study is in the proposal of a scalable and adaptable framework for discovering a thorough methodology to utilize for computing similarity between people or products.

REFERENCES

al., A.-B. E., 2019. A Study on the Accuracy of Prediction in Recommendation System Based on Similarity Measures. *Baghdad Science Journal*.

Atashkar, M., Safi-Esfahani, F., 2020. Item-Based Recommender Systems Applying Social-Economic Indicators. *SN Computer Science*, 1(2), 113. https://doi.org/10.1007/s42979-020-0115-8.

Fkih, F., 2021. Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison. *Journal of King Saud University - Computer and Information Sciences*.

Gazdar, A., Hidri, L., 2020. A new similarity measure for collaborative filtering based recommender systems. *Knowl. Based Syst.*, 188.

Khojamli, H., Razmara, J., 2021. Survey of similarity functions on neighborhood-based collaborative filtering. *Expert Syst. Appl.*, 185, 115482.

Nudrat, S., Khan, H. U., Iqbal, S., Talha, M. M., Alarfaj, F. K., Almusallam, N., 2022. Users Rating Predictions Using Collaborating Filtering Based on Users and Items Similarity Measures. *Computational Intelligence and Neuroscience*, 2022, 2347641. https://doi.org/10.1155/2022/2347641.

Singh, P. K., Sinha, M., Das, S., Choudhury, P., 2020. Enhancing recommendation accuracy of item-based collaborative filtering using Bhattacharyya coefficient and most similar item. *Applied Intelligence*, 50(12), 4708–4731. https://doi.org/10.1007/s10489-020-01775-4.

Stitini, O., Kaloun, S., Bencharef, O., 2020. Latest Trends in Recommender Systems applied in the medical domain: A Systematic Review. *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*.

Stitini, O., Kaloun, S., Bencharef, O., 2021. The recommendation of a practical guide for doctoral students using recommendation system algorithms in the education field.

Stitini, O., Kaloun, S., Bencharef, O., 2022. An Improved Recommender System Solution to Mitigate the Over-Specialization Problem Using Genetic Algorithms. *Electronics*, 11(2). https://www.mdpi.com/2079-9292/11/2/242.

Suganeshwari, G., Ibrahim, S., 2018. A comparison study on similarity measures in collaborative filtering algorithms for movie recommendation.

Zubair, S., Sabri, M. A., Khan, A., 2019. Correlation among similarity measurements for collaborative filtering techniques: An improved similarity metric.

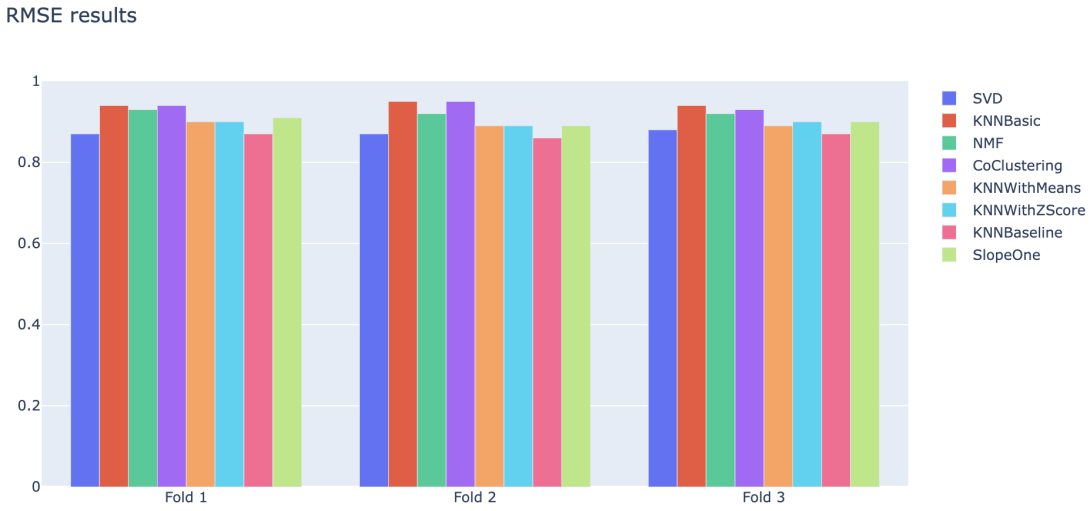

Figure 8. RMSE results using Model-based collaborative filtering.

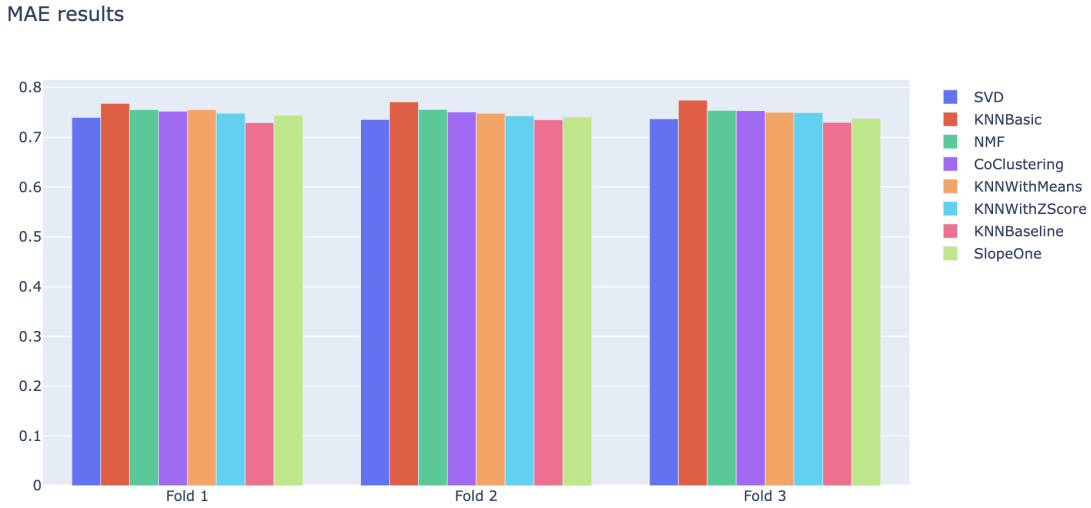

Figure 9. MAE results using Model-based collaborative filtering.

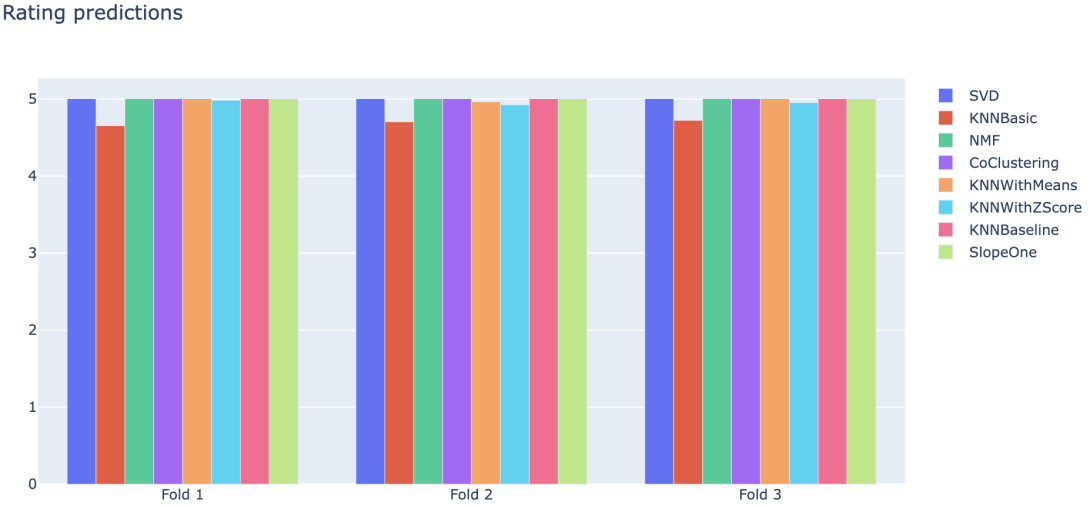

Figure 10. Rating prediction results using all model-based recommendation algorithms.