International Journal of Computing Sciences Research (ISSN print: 2546-0552; ISSN online: 2546-115X)
Vol. 8, pp. 3070-3087
doi: 10.25147/ijcsr.2017.001.1.204
https://stepacademic.net

Short Paper*

An Enhanced Content-based Filtering Using Maximal Marginal Relevance

Samantha Gwyn M. Aranzamendez
Computer Science Department, Pamantasan ng Lungsod ng Maynila (University of the City of Manila), Manila, Philippines
sgmaranzamendez2020@plm.edu.ph
(corresponding author)

Joshua Caleb D. Bolito
Computer Science Department, Pamantasan ng Lungsod ng Maynila (University of the City of Manila), Manila, Philippines
jcdbolito2020@plm.edu.ph

Aron Christoper R. Rafe
Computer Science Department, Pamantasan ng Lungsod ng Maynila (University of the City of Manila), Manila, Philippines
acrrafe2020@plm.edu.ph

Jamillah S. Guialil
Computer Science Department, Pamantasan ng Lungsod ng Maynila (University of the City of Manila), Manila, Philippines
jsguialil@plm.edu.ph

Dan Michael A. Cortez
Computer Science Department, Pamantasan ng Lungsod ng Maynila (University of the City of Manila), Manila, Philippines
dmacortez@plm.edu.ph

Raymund M. Dioses
Computer Science Department, Pamantasan ng Lungsod ng Maynila (University of the City of Manila), Manila, Philippines
rmdioses@plm.edu.ph

Date received: May 7, 2024
Date received in revised form: July 13, 2024
Date accepted: July 13, 2024

Recommended citation:

Aranzamendez, S. G. M., Bolito, J. C. D., Rafe, A. C. R., Guialil, J. S., Cortez, D. M. A., & Dioses, R. M. (2024). An Enhanced Content-based Filtering Using Maximal Marginal Relevance. *International Journal of Computing Sciences Research, 8*, 3070-3087. https://doi.org/10.25147/ijcsr.2017.001.1.204

**Special Issue on International Research Conference on Computer Engineering and Technology Education (IRCCETE). Guest Associate Editors: Dr. Roben A. Juanatas (National University-Manila) and Dr. Nelson C. Rodelas (University of East).*

Abstract

Purpose – The study aims to enhance Content-based Filtering by diversifying its recommended items to combat overspecialization. It traditionally recommends items that are directly related to the user profile, preventing users from discovering newer sets of items.

Method – Maximal Marginal Relevance is integrated into the algorithm – a re-ranking algorithm, developed by Carbonell and Goldstein that enhances the diversity of items retrieved by information retrieval systems – to enhance Content-based Filtering and address the underlying overspecialization problem.

Results – By integrating Maximal Marginal Relevance, the modified algorithm addressed overspecialization. Out of all the tested values of lambda (λ) for MMR, the enhanced Content-based Filtering (CBF-MMR) with λ = 0.7 showed the most prominence, having a good balance between relevance and diversity of recommendations. On average, it improved upon the original algorithm by 48.51% in Precision, 6.40% in Recall, 28.12% in F-Score, and 275.45% in Diversity.

Conclusion – Results show that integrating Maximal Marginal Relevance to Content-based Filtering (CBF-MMR) improves the diversity of recommendations. Due to the re-ranking process added by the Maximal Marginal Relevance, the average Precision, Recall, and F-Score also improved.

Recommendations – The authors of this study suggest further work on Content-based Filtering with faster re-ranking algorithms, application of the enhanced algorithm to other larger datasets such as GroupLens' MovieLens 10M dataset, application of the enhanced algorithm to a different domain, and enhancement of the Maximal Marginal Relevance algorithm to be applied in Content-based Filtering.

Research Implications – The successful integration of Maximal Marginal Relevance (MMR) in a Content-based Filtering algorithm opens new possibilities for enhancing the diversity and relevance of recommendations of various types of recommender systems.

Practical Implications – Beyond the movie recommender system this study was applied to, this study has profound practical implications on other domains that utilize recommender systems including but not limited to the domains of entertainment, e-commerce, and information retrieval platforms.

Keywords – recommender system, content-based filtering, maximal marginal relevance, overspecialization

INTRODUCTION

Data has grown exponentially through the streams of sites, systems, and applications. Regardless of the platform, when consumers are faced with an abundance of massive quantities of data, it complicates the process of choosing the best and/or most appropriate items for them. With the rise in the number of digital services directed at consumers in various domains such as shopping, music, movies, travel, and articles, recommender systems have become pivotal in helping users navigate relevant content and products (Al-Ghuribi & Noah, 2021). Such digital services use recommender systems to simplify the consumers' overall user experience.

Recommender systems consist of an algorithm that is intended to make recommendations. An algorithm follows a specific set of instructions, that requires calculations, to solve a problem. The increased use of recommender systems across the Internet and digital media has helped users by facilitating their decision-making. By utilizing the data provided by the user, recommender systems can make suggestions according to the user's needs and preferences, making a relevant and more personalized experience (Ebadi & Krzyzak, 2016). Among the prominent platforms that utilize recommender systems include Amazon – an e-commerce platform that utilizes user activity to recommend products a user might want to buy (Amazon, n.d.); Netflix – a streaming platform that utilizes user activity to recommend movies or shows a user might enjoy (Netflix, n.d.); and Facebook – a social media platform that utilizes recommender systems to help its users discover new communities and content (Facebook, n.d.).

Among the types of recommender systems is Content-based Filtering which provides recommendations based on the description of an item a user interacted with (Manjula, 2016; Al-Bashir et al., 2017). This systematic approach is built on things that bear a significant correlation to a user's profile. In a recommender system, a user profile is created from the given items that were favored by the user. User profiles consist of terms or features that the user had previously preferred and rated. These accumulated data items are collected and grouped into various item profiles, depending on their features or descriptions. Content-based filtering recommends items that are most related to the user profile. Recommendations are the item or feature suggestions given by the recommender system that are correlated to the user's profile. This creates a personalized approach to creating suggestions that cater to the user's preferences.

Despite its relevant functionality, Content-based Filtering is riddled with limitations as well. Among these limitations of the recommendations provided by Content-based Filtering is overspecialization where the algorithm recommends those items that are directly related to the user profile based on only a few attributes (Son & Kim, 2017), limiting the system to only recommend items that have a high similarity score compared to the user profile which rules out those newer sets of items from being recommended (Stitini et al., 2023). Hence, if a user loves and has only interacted with comedy films, the algorithm will not recommend films from a different genre since they only interacted with comedy films (Saat et al., 2018). This is further supported by the qualitative results of the study of Lokesh (2019) where out of the 18 genres in the dataset, given a movie name, the top 20 movies recommended by the algorithm are limited only to the genres of the given movie name.

The overspecialization problem is one of the most common problems of recommender systems as the algorithm rules out those newer item groups of items or features (Stitini et al., 2023). This means that the recommendations and suggestions given by the Content-based Filtering recommender system struggle to give an item or feature suggestions without any prior ratings or reviews made. This equates to giving out limited and biased item recommendations. This greatly affects the effectiveness of traditional Content-based Filtering and its algorithm in generating a relevant and diverse item recommendation.

In this study, an enhanced Content-based Filtering will be developed that addresses the issue of overspecialization and improves the diversity of the recommended items by using Maximal Marginal Relevance – a re-ranking algorithm developed in 1998 by Carbonell and Goldstein to enhance the diversity of items retrieved by information retrieval systems. The enhanced algorithm (CBF-MMR) will then be tested using GroupLens' MovieLens 1M Dataset; evaluated using the metrics of Precision, Recall, F-Score, and Diversity; and compared against the original algorithm and the enhancements by the studies of Stitini et al. (2022) and Cordero et al. (2022) to gauge how it ranks against other variants.

LITERATURE REVIEW

Content-based Filtering

Content-based Filtering is a type of recommender system that uses the features of items and the preferences of users to suggest items that are like what the user liked before. A content-based recommender suggests items by using the data provided by the users in the form of ratings or other forms of interaction (Aziz & Fayyaz, 2021).

At the basic level, Content-based Filtering relies on two sources of data: the item description and user profile which is generated from actions by the user towards various items or information provided by the user (Saat et al., 2018). Figure 1 shows the basic process of how Content-based Filtering works.

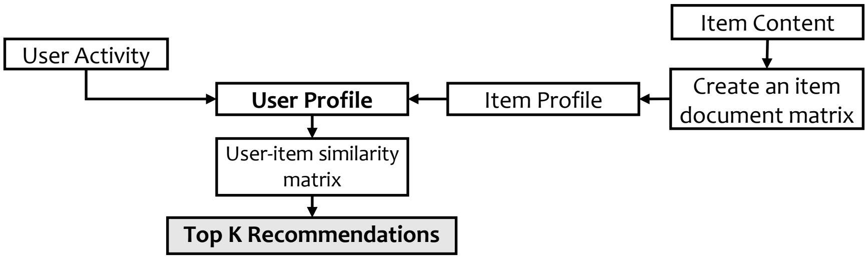

Figure 1. The basic process of Content-based Filtering

Overspecialization

Due to the user and feature-centered nature of Content-based Filtering, several drawbacks arise from the said algorithm. Among these drawbacks is the phenomenon called overspecialization. The process occurs when the algorithm recommends those items that are directly related to the user. However, these suggestions rule out those newer sets of items (Stitini et al., 2023) as it is more likely to recommend something that is like the user's profile (Saat et al., 2018). This poses a problem since when the given recommendations are closely related, the algorithm promotes the same types of items to users (Isinkaye et al., 2015; Thorat et al., 2015). Hence, the algorithm suffers from potentially poor diversity when it comes to generating recommendations (Stitini et al., 2023).

In addition to this, the system isn't prepared for anything unexpected, as it has a shortage of tools for investigating such phenomena. This means that new items with unrated features are less likely to be recommended (Stitini et al., 2022). Items similar to the user profile, are the ones being recommended first to the user. These biased results are recommended due to the computation of similarity scores that are repeatedly done and are only based on limited factors (Son & Kim, 2017).

Existing Enhancements on Content-based Filtering

The study conducted by Stitini et al. (2022) utilized Revolutionary Recommendation System Genetic Algorithms to provide flexible recommendations to its users. The proposed method considers all the available items in the recommendation list, instead of choosing specific items. The solution consisted of initializing the population and obtaining the fitness value of each item by getting the feature (genre) score of the item. The genetic algorithm is applied then it selects the optimal recommendation. The metrics consisted of measuring the diversity, novelty, precision, and recall of the proposed methods (using RRSGA) in comparison to a traditional content-based filtering algorithm. The results showed that the proposed algorithm performed better than the traditional algorithm, based on the generated diversity, novelty, precision, and recall scores.

The study conducted by Stitini et al. (2023) specified a general solution to all the given problems that are said to be within content-based filtering. Their proposed model called the Ideal Solution Mitigating Content Disadvantages based on Three Phases (ISMCD3P), uses NLP techniques, Popularity, and Metrics applications in each phase, respectively. It is then compared to traditional content-based filtering as well as to Content-based Filtering's existing enhancements. The metrics used to compare data were based on novelty, precision, and recall. The proposed model successfully solved the following issues; however, it is only limited since it utilized a single data set while testing. It was recommended to explore other ones to show the proposed model's effectiveness.

The approach of the study conducted by Son and Kim (2017) is to implement multi-attribute networks by using centrality and clustering techniques. The main idea is to do network analysis which analyzes the direct and indirect relationship between items to solve the sparsity and over-specialization problem of content-based filtering. The metrics used in this study are precision and recall. The obtained results are compared to other techniques like Traditional Content-based Filtering, Feature Weighting, and the use of Linked Open Data.

The modified Content-based Filtering approach by Cordero et al. (2022) integrates K-Nearest Neighbors (KNN) and Percentile Concept to resolve overspecialization. The method utilized KNN for items with relatively high cosine scores in within the specified trunk list. Hence it divided items that are not novel to the user and such items are not like one another. The percentile concept establishes a selection of values which a recommendation shall be coming from. The results show that both implemented methods have seemed to be effective as they showed transparently diverse results based on the percentile of an item in the cosine similarity matrix.

Maximal Marginal Relevance

Maximal Marginal Relevance was created to improve the diversity of the retrieved documents by Information Retrieval (IR) search engines (Carbonell & Goldstein, 1998). A document is of "high marginal relevance" if it is relevant to the query and exhibits minimal similarity to previously selected documents that are retrieved by the IR search engine. The formula for Maximal Marginal Relevance is seen in Equation 1 where *C* is the collection of documents; D_i are the individual documents in *C*; *Q* is the query; *R* is the relevant documents in *C*; *S* is the subset of documents in *R* that is already selected; *R\S* is the set difference or documents in *R* that are not yet in *S*; Sim_1 is the similarity metric used in document retrieval and relevance ranking between documents and a query; Sim_2 can be the same as Sim_1 or a different metric; and λ is a parameter which determines the accuracy or diversity ranking among the documents in *R* – the higher the value of λ, the better the relevance (1) and the lower the value of λ, the better the diversity (0). For optimal results, balance the value of λ according to preference.

$$MMR = Arg \max_{D_i \in R \setminus S} \left[\lambda(Sim_1(D_i, Q) - (1 - \lambda) \max_{D_i \in S} Sim_2(D_i, D_j) \right] \qquad \text{Equation 1}$$

In the study of Liu et al. (2020), Maximal Marginal Relevance was compared against other recommendation methods and shined in terms of diversity by being the second-best recommendation algorithm in the ML-100K dataset, the best recommendation algorithm in the ML-1M dataset (the dataset to be used in this study), and the second-best recommendation algorithm in Anime dataset. Maximal Marginal Relevance is also proven to be applicable in recommender systems and has been used to create a novel Recommendation Method that is focused on diversity (Luan et al., 2018).

Due to its simplicity and ability to diversify retrieved documents while considering relevance, Maximal Marginal Relevance will be useful for addressing the overspecialization issues faced by Content-based Filtering.

METHODOLOGY

This study integrates Maximal Marginal Relevance into Content-based Filtering to mitigate overspecialization and improve recommendation diversity. Performance assessment will involve evaluating Precision, Recall, F-Score, and Diversity metrics, comparing against the original algorithm and enhancements by Stitini et al. (2022) and Cordero et al. (2022) to gauge relative performance.

The diagram for the flow of the enhanced Content-based Filtering (CBF-MMR) can be seen in Figure 2.

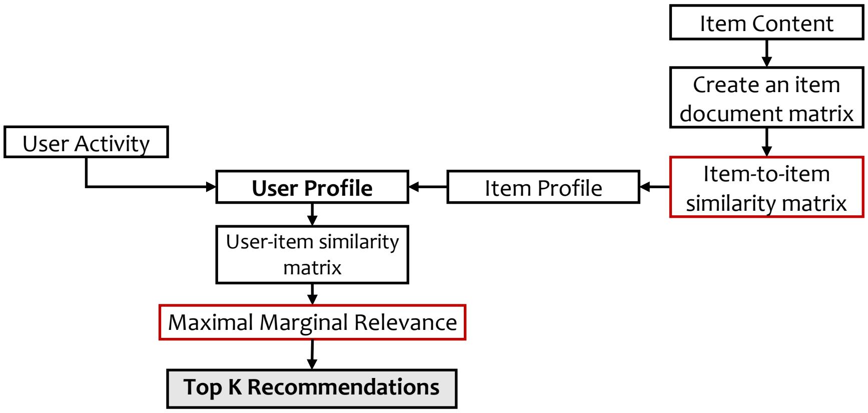
Item Content
Create an item document matrix
User Activity
User Profile
Item Profile
Item-to-item similarity matrix
User-item similarity matrix
Maximal Marginal Relevance
Top K Recommendations

Figure 2. Diagram of the enhancement on Content-based Filtering.

Maximal Marginal Relevance will be used on the user-item similarity matrix to compute relevance scores for each item. The diversity parameter (λ), ranging from 0 to 1, and the item-to-item similarity matrix, balance recommendation relevance and diversity. A λ closer to 0 yields more diverse but less relevant recommendations, while a λ closer to 1 prioritizes relevance over diversity.

Specifically, Maximal Marginal Relevance was integrated into the Content-based Filtering through the following steps:

1. Computation of Relevance Score
    a. Item Selection
        - Before computing the relevance score, the item is selected from the user-item similarity matrix.

    b. Relevance Score Calculation
        - Calculate the relevance score of the selected item based on its Term Frequency-Inverse Document Frequency (TF-IDF) score – a score that reflects the importance of a term (in this case, a movie genre) in a document (a movie). With the sum of the TF-IDF scores of the genres associated with the item, the relevance score is determined.

2. Calculation of MMR Score
    - Compute the MMR score of an item by combining its relevance score with its dissimilarity from already recommended items. This score is calculated using the diversity parameter (λ) and item-to-item similarity matrix, enabling adjustment between relevance and diversity. Higher λ values emphasize relevance, while lower values prioritize diversity. By subtracting the maximum similarity score of the item from already recommended items, the MMR score balances relevance and dissimilarity, aiding in selecting diverse yet relevant recommendations.

3. Repeat Selection Process
    - After computing MMR scores for each item, the selection iterates until the desired number of recommendations is reached. This process involves repeatedly calculating MMR scores for the remaining items and selecting the item with the highest score. Through iterative selection, the recommendation list is curated to provide diverse, yet relevant options aligned with the user's preferences or profile.

The original algorithm selects recommendations solely based on similarity to the user profile, favoring close matches. In contrast, introducing Maximal Marginal Relevance (MMR) adds diversity by considering the user profile alongside a diversity parameter (λ) and item-to-item similarity matrix. This allows fine-tuning of recommendation diversity: higher λ for more similar items, lower λ for less similarity.

Dataset

Maximal Marginal Relevance's impact on Content-based Filtering will be assessed using GroupLens' MovieLens 1M Dataset by Harper & Konstan (2015). This dataset includes 1,000,209 ratings of 3,900 movies across 18 genres by 6,040 users. Recommendations will be generated based on the genres of movies rated by the user. See Table 1 for a data summary.

Table 1. Specifications of MovieLens 1M Dataset

Properties	MovieLens Dataset
Number of users	6,040
Number of movies	3,900
Number of genres	18
Number of reviews/ratings	1,000,209

The 3,900 movies from the dataset are classified into one or several of the 18 different genres listed in Table 2.

Table 2. Movie genres of MovieLens 1M Dataset

Action	Children’s	Documentary	Film-Noir	Mystery	Thriller
Adventure	Comedy	Drama	Horror	Romance	War
Animation	Crime	Fantasy	Musical	Sci-Fi	Western

This dataset was specifically selected for this study due to how widely it has been used by other studies to gauge the performance of their enhancements on Content-based Filtering (Stitini et al., 2022; Lokesh, 2019; Son & Kim, 2017).

Metrics

Precision, Recall, and F-Score are commonly used as offline evaluation metrics for recommender systems, aiming to gauge the relevance of recommended items to users (Dalianis, 2018; Shani & Gunawardana, 2015). These metrics are also widely employed by other studies to assess system enhancements (Santos et al., 2014; Kundur et al., 2016; Shu et al., 2017; Bhagavatula et al., 2018). See Table 3 for the variables of the said metrics.

Table 3. Variables used in the equation of Precision, Recall, and F-Score

Variable		Description
True Positive	***tp***	Number of items retrieved that are expected to be liked by the user.
False Positive	***fp***	Number of items retrieved that are expected to not be liked by the user.
False Negative	***fn***	Several items that were expected to be liked by the user were not retrieved.

The definition, significance, and equation of Precision, Recall, and F-Score as a metric in recommender systems are as follows:

- **Precision** – Measures the fraction of correct instances retrieved among all retrieved instances. This metric determines the proportion of relevant items among the items recommended by the system. The formula for precision can be seen in Equation 2.

$$P = \frac{tp}{tp+fp} \qquad \text{Equation 2}$$

- **Recall** – Measures the fraction of correct instances retrieved among all correct instances that were and were not retrieved. This metric determines the proportion of relevant items that were recommended over the total amount of relevant items in the dataset. The formula for recall can be seen in Equation 3.
-

$$R = \frac{tp}{tp+fn} \qquad \text{Equation 3}$$

-
- **F Score** – Measures the balance between precision and recall. The formula for the F Score can be seen in Equation 4.
-

$$F\ Score = F1 = F = 2 * \frac{P*R}{P+R} \qquad \text{Equation 4}$$

The larger the Precision, Recall, and F Score (1 – best), the better the performance; the lower (0 – worst) it is, the worse it becomes.

Since this study aims to address the overspecialization issue of Content-based Filtering by diversifying the recommendations with Maximal Marginal Relevance, Diversity will be used as a metric to measure how different the recommended items are from each other (Stitini et al., 2022). The formula for diversity can be seen in Equation 5.

$$Diversity = 1 - Similarity \qquad \text{Equation 5}$$

Considering that the MovieLens dataset has rating data that ranges between one (lowest) to five (highest) points, to determine if a recommended movie would be liked by a user or not, the formula for the determining gauge can be seen in Equation 6 based on the study of Son & Kim (2017).

$$E = \sum_{c=min}^{max} c \times P(c) \qquad \text{Equation 6}$$

Where *E* serves as a user's border rating of whether they like a movie or not, *c* is the rating points, *P(c)* is the proportion of *c* for a given user, *min* is the minimum possible item rating, and *max* is the maximum possible item rating. If an item has a mean rating higher or equal to *E*, then the item is classified as preferred by the user. If an item has a mean rating lower than *E*, the item is classified as not preferred by the user.

To quantitatively evaluate the performance of the enhancement on Content-based Filtering, based on the genres of movies rated by the user, a set of Top K recommendations was given to 200 randomly selected users from the MovieLens 1M Dataset which was done in 10 iterations.

Precision, Recall, F-Score, and Diversity of the recommendations were measured and the results of the original Content-based Filtering, the enhanced algorithm by Stitini et al. (2022) with Genetic Algorithm parameters of 400 population and 50 generations with 5% mutation and 5% crossover probability, the enhanced algorithms by Cordero et al. (2022), and the enhancement of this study (CBF-MMR) at varying values of lambda were compared to determine if this study's enhancement is better and if so, which lambda (λ) has the best recommendations in terms of the given metrics.

Algorithms will be tested across different recommendation counts (K) to determine their performance. Average results across all K values will be compared to identify the best-performing algorithm, with the top results highlighted in bold and the second-best results underlined in the tables.

RESULTS

Table 4 compares Precision among algorithm versions across recommendation counts. CBF-MMR with λ = 0.7 ranks best for K = 3, K = 5, and K = 11, and second-best for K = 9. Following, CBF-MMR with λ = 0.5 ranks best for K = 7 and K = 9, and second-best for K = 3 and K = 11. Their average precision across different K values further supports this, with λ = 0.7 leading and λ = 0.5 following closely.

Table 4. Comparison of Precision

K	***CONTENT-BASED FILTERING VARIANTS***							
	Original	***Stitini et al. (2022)***	***Cordero et al. (2022)***		***CBF-MMR***			
			K-Nearest Neighbors	***Percentile Concept***	***λ = 0***	***λ = 0.5***	***λ = 0.7***	***λ = 1***
1	0.276	0.3458	**0.3785**	0.3095	0.1643	0.285	0.29	0.29
3	0.2915	0.3326	0.3121	0.2928	0.2428	0.5142	**0.5268**	0.3721
5	0.2974	0.3361	0.3134	0.2992	0.3766	0.3582	**0.4848**	0.4222
7	0.3069	0.3319	0.3088	0.2919	0.4159	**0.4473**	0.4088	0.4447
9	0.3052	0.3304	0.3051	0.2972	0.433	**0.4667**	0.4435	0.3886
11	0.3024	0.3226	0.3061	0.2966	0.4054	0.4269	**0.4886**	0.4145
AVE	0.2966	0.3332	0.3207	0.2979	0.3397	0.4164	**0.4404**	0.3887

Table 5 compares Recall among algorithm variants across recommendation counts. CBF-MMR with λ = 0.5 ranks best for K = 3, K = 7, and K = 9, and second-best for K = 11. This means λ = 0.5 has the best average proportion of relevant items that were recommended over the total amount of relevant items in the dataset.

Table 5. Comparison of Recall

K	***CONTENT-BASED FILTERING VARIANTS***							
	Original	***Stitini et al. (2022)***	***Cordero et al. (2022)***		***CBF-MMR***			
			K-Nearest Neighbors	***Percentile Concept***	***λ = 0***	***λ = 0.5***	***λ = 0.7***	***λ = 1***
1	0.0666	0.0865	**0.1279**	0.0766	0.0225	0.027	0.033	0.0344
3	0.2086	0.2145	0.2579	0.2019	0.1292	**0.2733**	0.2481	0.1937
5	0.3403	0.3614	**0.4344**	0.3498	0.3401	0.3182	0.3935	0.3309
7	0.5196	0.5071	0.5641	0.5005	0.5604	**0.5987**	0.4543	0.58
9	0.6696	0.6545	0.6627	0.6485	0.7708	**0.8111**	0.7122	0.6329
11	0.8154	0.7891	0.7763	0.7847	0.8562	0.8667	**0.9468**	0.8322
AVE	0.4367	0.4355	0.4706	0.4270	0.4465	**0.4825**	0.4647	0.4340

Table 6 compares F-Scores of various algorithm variants across recommendation counts. CBF-MMR with λ = 0.7 ranks best for K = 5 and K = 11, and second-best for K = 3. Despite λ = 0.5 having more top rankings, the average F-Score across all K values favors λ = 0.7, indicating a better average balance of Precision and Recall.

Table 6. Comparison of F-Score

K	***CONTENT-BASED FILTERING VARIANTS***							
	Original	***Stitini et al. (2022)***	***Cordero et al. (2022)***		***CBF-MMR***			
			K-Nearest Neighbors	***Percentile Concept***	***λ = 0***	***λ = 0.5***	***λ = 0.7***	***λ = 1***
1	0.1073	0.1384	**0.1912**	0.1228	0.0396	0.0493	0.0593	0.0615
3	0.2439	0.2608	0.2824	0.2390	0.1687	**0.3569**	0.3373	0.2548
5	0.3174	0.3483	0.3641	0.3225	0.3574	0.3370	**0.4344**	0.3710
7	0.3859	0.4012	0.3991	0.3687	0.4775	**0.5120**	0.4304	0.5034
9	0.4193	0.4391	0.4178	0.4076	0.5545	**0.5925**	0.5466	0.4815
11	0.4412	0.4580	0.4391	0.4305	0.5503	0.5720	**0.6446**	0.5534
AVE	0.3190	0.3410	0.3490	0.3152	0.3580	0.4033	**0.4088**	0.3709

Table 7 compares Diversity among algorithm versions across recommendation counts. CBF-MMR with λ = 0 ranks best for all K values except K = 1, followed by λ = 0.7, which ranks second-best for K = 5 and K = 7. The average diversity across different K values reinforces this, with λ = 0 leading and λ = 0.7 following closely.

Table 7. Comparison of Diversity

K	***CONTENT-BASED FILTERING VARIANTS***							
	Original	***Stitini et al. (2022)***	***Cordero et al. (2022)***		***CBF-MMR***			
			K-Nearest Neighbors	***Percentile Concept***	***λ = 0***	***λ = 0.5***	***λ = 0.7***	***λ = 1***
1	0.185	0.5797	0.5852	0.3907	0.5867	**0.7097**	0.7057	0.7062
3	0.1867	0.5815	0.5471	0.3897	**0.7659**	0.7426	0.7297	0.6913
5	0.1879	0.5865	0.5578	0.392	**0.8337**	0.698	0.7155	0.6957
7	0.1887	0.5867	0.559	0.39	**0.8741**	0.6682	0.7192	0.6863
9	0.1893	0.5865	0.5527	0.3884	**0.8986**	0.6706	0.6928	0.693
11	0.1898	0.5885	0.5536	0.3897	**0.8735**	0.6787	0.6699	0.6955
AVE	0.1879	0.3410	0.3490	0.3152	**0.8054**	0.6946	0.7055	0.6947

DISCUSSION

Overall, integrating Maximal Marginal Relevance into Content-based Filtering enhanced Precision, Recall, F-Score, and Diversity. Notably, CBF-MMR with λ = 0.7 achieves the best balance between relevance and diversity, showcasing superior average Precision and Recall as seen in F-Score results (Table 6) and the second-best assortment of recommendations as seen in Diversity results (Table 7).

To better visualize how much CBF-MMR with λ = 0.7 improved upon the original algorithm, the percentage of improvement can be acquired as seen in Equation 7, where given a metric, *E* is the value of the enhanced algorithm and *O* is the value of the original algorithm.

$$\% \, Improvement = \left(\frac{E-O}{O}\right) * 100 \qquad \textit{Equation 7}$$

Comparing the average results yielded from the different metrics by the CBF-MMR with λ = 0.7 and the original algorithm, it can be computed that the enhancement improved upon the original algorithm by 48.51% in terms of Precision, 6.40% in terms of Recall, 28.12% in terms of F-Score, and 275.45% in terms of Diversity.

CONCLUSIONS AND RECOMMENDATIONS

The integration of Maximal Marginal Relevance (MMR) in Content-based Filtering (CBF) enhances recommendation diversity, as evident in Table 7. Even with λ = 1, which prioritizes fewer diverse recommendations, CBF-MMR improves diversity over two-fold compared to the original algorithm. This addresses overspecialization, where the algorithm tends to recommend only closely related items, hindering users from discovering new content. Additionally, MMR's re-ranking process enhances average Precision, Recall, and F-Score, as shown in Tables 4–6.

Of all the tested values of lambda, CBF-MMR with λ = 0.7 showed the most prominence with its good balance between relevance and diversity of recommendations. Specifically, CBF-MMR with λ = 0.7 has the best average balance between Precision and Recall as seen in the results of F-Score in Table 6; and has the second-best average assortment of recommendations as seen in the results of Diversity in Table 7.

On average, the CBF-MMR with λ = 0.7 improved upon the original algorithm by 48.51% in terms of Precision, 6.40% in terms of Recall, 28.12% in terms of F-Score, and 275.45% in terms of Diversity.

For future studies, find a better re-ranking algorithm than Maximal Marginal Relevance that requires less computational time. Apply the enhanced algorithm to other larger datasets such as GroupLens' MovieLens 10M dataset or apply it to a dataset of a different domain.

IMPLICATIONS

The integration of Maximal Marginal Relevance (MMR) to address the overspecialization issue that Content-based Filtering suffers from opens new possibilities for enhancing the diversity and relevance of recommendations. This has profound implications for various domains including entertainment, e-commerce, and information retrieval platforms. Future research endeavors can further improve the algorithm by addressing other weaknesses of Content-based Filtering and exploring other possible domain of integration that hopefully encompasses one or several of the 17 United Nations Sustainable Development Goals for a broader impact.

ACKNOWLEDGEMENT

The authors would like to thank Pamantasan ng Lungsod ng Maynila's College of Information System and Technology Management and Computer Science Department for continuous encouragement. The researchers would like to express sincere appreciation to the Department of Science and Technology's NCR Regional Office and Science Education Institute for their assistance. Lastly, friends and family for their unwavering support.

FUNDING

This study is partially funded by the Department of Science and Technology's NCR Regional Office and Science Education Institute, specifically supporting incurred fees throughout the duration of the study. The remaining costs were covered by the researchers themselves, using their funds.

DECLARATIONS

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the study.

Informed Consent

Due to the nature of the study being an enhancement of an algorithm, the creation of this study did not involve any human participants. Hence, informed consent was not applicable.

Ethics Approval

This study did not involve human or animal subjects. Therefore, did not require ethics approval, making it not applicable to this study.

REFERENCES

Al-bashiri, H., Abdulhak, M. A., Romli, A., & Hujainah, F. (2017). Collaborative filtering recommender system: Overview and challenges. *Advanced Science Letters, 23*(9), 9045-9049(5). https://doi.org/10.1166/asl.2017.10020

Al-Ghuribi, S. M., & Noah, S. A. (2021, September). *A comprehensive overview of the recommender system and sentiment analysis*. Retrieved October 3, 2023, from https://www.researchgate.net/publication/318534250_A_hybrid_multi-criteria_hotel_recommender_system_using_explicit_and_implicit_feedbacks

Amazon. (n.d.). *Recommendations*. Retrieved April 4, 2024, from Amazon: https://www.amazon.com/gp/help/customer/display.html?nodeId=GE4KRSZ4KAZZB4BV

Bhagavatula, C., Feldman, S., Power, R., & Ammar, W. (2018, June). Content-based citation recommendation. In M. Walker, H. Ji, & A. Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1022

Carbonell, J., & Goldestein, J. (1998, August 1). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Melbourne, Australia: Association for Computing Machinery. https://doi.org/https://doi.org/10.1145/290941.291025

Cordero, N. J., Canlas, J. C., Mata, K. E., Regala, R. C., Blanco, M. C., Cortez, D. M., & Alipio, A. J. (2022, May 5). Modified Content-based filtering method using K-Nearest Neighbors and Percentile Concept. *International Journal of Research Publications, 100*(1), 20-33. https://doi.org/10.47119/IJRP1001001520223119

Dalianis, H. (2018, May 15). Evaluation metrics and evaluation. In *Clinical Text Mining* (pp 45-53). Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-78503-5_6

Ebadi, A., & Krzyzak, A. (2016, January). A hybrid multi-criteria hotel recommender system using explicit and implicit feedbacks. *International Journal of Computer and Information Engineering*, *10*(8), 1450-1458.

Facebook. (n.d.). *What are recommendations on Facebook?* Retrieved April 4, 2024, from Facebook: https://www.facebook.com/help/1257205004624246/?helpref=uf_share

Harper, F. M., & Konstan, J. A. (2015). The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems,* 5(4), 1-19. https://doi.org/https://doi.org/10.1145/2827872

Isinkaye, F. O., Folajimi, Y., & Ojokoh, B. (2015). Recommendation systems: Principles, methods, and evaluation. *Egyptian Informatics Journal 16(3)*, 261-273. https://doi.org/10.1016/j.eij.2015.06.005

Kundur, N. C., Dhulavvagol, P. M., & Prasad, M. R. (2016). Recommendation system based on content filtering for specific commodity. *International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS), V*(VII). Retrieved November 11, 2023, from https://www.ijltemas.in/DigitalLibrary/Vol.5Issue7/25-29.pdf

Liu, Y., Xiao, Y., Wi, Q., Miao, C., Zhang, J., Zhao, B., & Tang, H. (2020). Diversified interactive recommendation with implicit feedback. *Proceedings of the AAAI Conference on Artificial Intelligence,* *34*(04), 4932-4939. https://doi.org/https://doi.org/10.1609/aaai.v34i04.5931

Lokesh, A. (2019). *A comparative study of recommendation systems* (unpublished manuscript). Retrieved October 10, 2023, from https://digitalcommons.wku.edu/theses/3166/

Luan, W., Liu, G., Jiang, C., & Zhou, M. (2018). MPTR: A Maximal Marginal Relevance-based personalized trip recommendation method. *IEEE Transactions on Intelligent Transportation Systems, 19*(11), 3461-3474. https://doi.org/10.1109/TITS.2017.2781138

Manjula, R. (2016). Content-based filtering techniques in recommendation system using user preferences. *7. International Journal of Innovations in Engineering and Technology (IJIET), 74,* 151. Retrieved October 3, 2023, from https://ijiet.com/wp-content/uploads/2016/12/20.pdf

Netflix. (n.d.). *How Netflix's recommendations system works*. Retrieved April 4, 2024, from Netflix: https://help.netflix.com/en/node/100639?q=recommend

Saat, N. I., Noah, S. A., & Mohd, M. (2018). Towards serendipity for content-based recommender systems. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(4-2), 1762-1769. https://doi.org/https://doi.org/10.18517/ijaseit.8.4-2.6807

Santos, I., Miñambres-Marcos, I., Carlos, L., Galán-García, P., Santamaría-Ibirika, A., & Bringas, P. G. (2014). Twitter content-based spam filtering. *Advances in Intelligent System- and Computing, Volume 239*. Springer International Publishing Switzerland. https://doi.org/https://doi.org/10.1007/978-3-319-01854-6_46

Shani, G., & Gunawardana, A. (2015). Evaluating recommender systems. In F. Ricci, L. Rokach, & B. Shapira, (eds.), *Recommender Systems Handbook* (pp. 265-308). https://doi.org/https://doi.org/10.1007/978-1-4899-7637-6_8

Shu, J., Shen, X., Liu, H., Yi, B., & Zhang, Z. (2017). A content-based recommendation algorithm for learning resources. *Multimedia Systems, 24,* 163-173. https://doi.org/https://doi.org/10.1007/s00530-017-0539-8

Son, J., & Kim, S. (2017). Content-based filtering for recommendation systems using multiattribute networks. *Expert Systems with Applications, 89*, 404-412. https://doi.org/10.1016/j.eswa.2017.08.008

Stitini, O., Kaloun, S., & Bencharef, O. (2022). An improved recommender system solution to mitigate the over-specialization problem using Genetic Algorithms. *Electronics*(11), 1-22. https://doi.org/ https://doi.org/10.3390/electronics11020242

Stitini, O., Kaloun, S., & Bencharef, O. (2023). Towards a robust solution to mitigate all content-based filtering drawbacks within a recommendation system. *International Journal of Systematic Innovation,* 7(7), 89-111. https://doi.org/10.6977/IJoSI.202309_7(7).0006

Thorat, P. B., Goudar, R. M., & Barve, S. (2015). Survey on Collaborative Filtering, Content-based Filtering, and Hybrid Recommendation System. *International Journal of Computer Applications, 110*(4), 31-36.

Author’s Biography

Samantha Gwyn Aranzamendez is a dedicated Computer Science major from Pamantasan ng Lungsod ng Maynila. She has a passionate interest in system analytics, SEO optimization, WordPress development, and software quality assurance. With a sharpened eye for innovation and a drive for continuous learning, she is ready to make major contributions to the constantly evolving scene of technology.

Joshua Caleb D. Bolito is a graduating computer science student of Pamantasan ng Lungsod ng Maynila, where he is pursuing a bachelor’s degree. Joshua is set to graduate in September 2024 with Magna Cum Laude honors As a skilled Software Developer, Joshua has honed his expertise in Frontend Development, specializing in the use of cutting-edge technologies such as React and Next.js.

Aron Christoper Rafe, is a fourth-year student pursuing a bachelor's degree in computer science at Pamantasan ng Lungsod ng Maynila. He's passionate about creating software solutions to benefit the community. He continually hones his soft and technical skills in his free time. He aspires to become an Android developer. Upon graduation, he looks forward to applying his growing skills and knowledge in mobile technology to be able to contribute to the community as an Android Developer.

Jamillah Guialil earned her Bachelor of Science in Computer Studies with a major in Computer Science (BSCS-CS) in 2018. Currently, she is pursuing her Master of Information

Technology (MIT) degree at Pamantasan ng Lungsod ng Maynila. In addition to her studies, Jamillah is working as a part-time faculty member at the College of Information Systems and Technology Management (CISTM) within Pamantasan ng Lungsod ng Maynila.

Dr. Dan Michael A. Cortez is the Vice President for Research, Academic, and Extension Services at Pamantasan ng Lungsod ng Maynila, with 10 years of teaching experience. He holds a Bachelor of Science in Information Technology and a Master of Science in Information and Communications Technology from the same university. He earned his Doctorate in Information Technology from the Technological Institute of the Philippines-Quezon City Campus. Dr. Cortez is a member of PSITE-NCR and the Computing Society of the Philippines, with research interests focused on cryptography and Data Mining, having authored multiple books and published research both locally and internationally.

Raymund M. Dioses is currently an Assistant Professor I at Pamantasan ng Lungsod ng Maynila, where he also chairs the Computer Science Department. He previously worked at CORE Gateway College Inc. for 8 years as a College Faculty and Chairperson of the Computer Education Department, and 5 years as a Teacher II at the Department of Education. He holds a Bachelor of Science in Computer Science from St. Jude College and a Master of Arts in Education majoring in Educational Management from CORE Gateway College. He is currently pursuing a Master of Information Technology majoring in Computer Education at Nueva Ecija University of Science and Technology.