

Hybrid-based Research Article Recommender System

Su-Anne Teh¹, Su-Cheng Haw^{2*}, Heru Agus Santoso³

^{1,2}*Faculty of Computing and Informatics, Multimedia University, 63100, Cyberjaya, Malaysia; E-mail: sucheng@mmu.edu.my*

³*Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia*

Abstracts: A recommender system, which might assist in providing clients with new information and a better experience, is becoming increasingly popular in this era of modernization. Recommender systems are often used by various platforms to provide new products to consumers, which may also help in improving product sales. Additionally, the recommender system is essential in academic domains. It is common for users to take a while to find and access the materials they need. The recommender system is now available, which could reduce the time spent looking for materials and improve student achievement. Therefore, it is crucial to explore more on the theory and implementation of the recommender system. This paper aims to study a few types of recommender system techniques and implement it in the research article recommender system. Additionally, related research on each of the three recommender systems will be reviewed, along with a description of the related study, the dataset used, and the evaluation method.

Keywords: Research Article Recommender System, Recommender System, Content-Based Filtering, Collaborative Filtering, Hybrid Filtering.

1. INTRODUCTION

Technology is advancing at an ever-increasing rate in this era of modernization, with everyone owning at least one digital device. Thanks to technological advancements, many things have been converted to digital form, and people today depend more on electronics. Traditional retailers have extensively used recommender systems, a type of business intelligence technology, to boost brand competitiveness in the e-commerce industry (Zhou et al., 2022). As recommender systems become increasingly prevalent in users' daily lives, recommendations are beginning to influence more decisions (Zaveri et al., 2023). Users can make decisions more quickly and accurately with a recommender system since the user will be presented with pertinent choices.

Today's average internet user does not want to spend time searching for a certain item and assumes the system will handle everything and provide them with efficient ideas. Conversely, online retailers are eager to understand their customers' interests to win them over as lifetime customers. If a trader knows their client's preferences, they will always have an advantage over other dealers or competitors. Decision-support technologies called recommender systems to use advanced algorithms to assist users in discovering less-known but potentially intriguing topics (Elahi et al., 2021).

Online sales frequently use recommender systems, giving customers access to new information and purchasing options (Zhou et al., 2022; Chew et al., 2020). For instance, the recommender systems in Shopee and Lazada will propose products that the customer would be interested in buying. By doing so, this might assist in increasing product sales in the platforms that were using recommender system. Recommender systems' primary duties often involve directing users to further information that they might be interested in or filtering incoming sources of data in accordance with their preferences (Karimi et al., 2018). Other than the recommender system that was utilised in e-commerce, recommender systems also play an important role in academic fields. According to (Khademizadeh et al., 2022), academic libraries deal with a sizable and expanding volume of data and a wide range of reading materials. Users frequently take a long time to look for and obtain the required resources. Academic libraries also play a crucial part in pupils' education as it might help raise students' performance.

2. LITERATURE REVIEW

2.1. Overview on Recommender System

Everyone nowadays owns at least one digital gadget, which tends to make life easier in various ways. Compared to the earlier decades of society where people interacted more with one another face to face, our lives now depend more on digital products due to advancements in science and technology. The third industrial revolution, known as digitization, is thought to have started in the later half of the 20th century. Undeniably, the development of digital products has greatly aided daily human life in areas such as entertainment, online shopping, banking, hotel and ticket booking, and others. However, as there are more products available for the user to select, which may cause information overload where the consumer finds it difficult to decide on what they want. Thus, this leads to the recommender system, which provides suggestions for products that the user might be interested in.

Recommender system provides recommendations to users for products that might interest them. Recommender systems are a tool that assists users in locating information of their interest and it typically analyses user preferences and makes suggestions for relevant targets. In order to develop theories of user and item affinities that can be utilised to identify well-matched pairings, various recommender systems analyse various data sources (Melville & Sindhvani, 2017). The recommender system's primary purpose is to propose the most relevant material for users, but it will also work as a filtering mechanism to reduce information overload (Joe & Raj, 2021, Isinkaye et al., 2015, Falk, 2019).

According to (Gulzar et al., 2018), there are four recommender systems: content-based, collaborative filtering, knowledge-based, and hybrid. A collaborative filtering recommender system examines previous interactions and makes product recommendations based on user ratings with similar users. Besides, recommender systems with content-based filtering techniques provide recommendations for items based on user profiles. Knowledge-based recommender systems use artificial intelligence algorithms to compare user and item similarities. These methods make advantage of comprehensive information about item features rather than depending on human evaluations. Combining collaborative filtering with content-based filtering was recommended as a hybrid approach to merging the standard recommender systems so it could lessen the problems with separate recommender systems.

Many industries heavily rely on recommender systems, including healthcare, e-learning, banking, marketing, and e-commerce. For instance, the e-commerce industry extensively used recommender systems, which helped them increase their consumer base and financial success. In the healthcare, it aims to supply its user (patient) with medical information that is meant to be highly relevant and tailored to an individual's need.

Efficient models are needed to provide accurate research recommendations due to the influx growth of published articles. Digital libraries are overloaded, making it time-consuming to find relevant paper. Literature search takes time, as such we need intelligent systems to make it to suggest for pertinent articles among millions available (Sharma et al., 2023).

2.2. Recommender System Techniques

Three basic types of filtering approaches can be used in recommender systems: collaborative filtering, content-based filtering, and hybrid filtering. The organisation may make it easier for the user to receive a trustworthy suggestion by implementing the right and appropriate techniques in the recommender system. Fig. 1 shows the techniques that can be used to implement a recommender system.

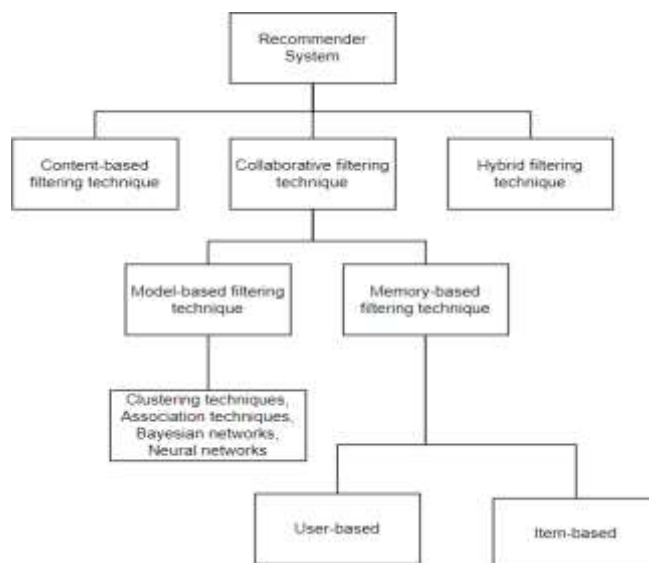


Fig. 1. Techniques Used in Recommender System

2.2.1. Content-Based Filtering Technique

The purpose of content-based filtering is to analyse the characteristics of things and recommend information that is related to what the user was interested in during the previous activity. This method determines the identical things that most closely resemble the user profile by comparing each item's properties with the user profile. In content-based filtering technique, each user normally will have a profile which includes all the relevant user information that could help gather the user's personal information and user characteristics such as name, gender, age, location, area of interest and so on. The process of collecting and identifying keyword-based data to create a structured profile and then visualising the knowledge gleaned from these results is known as user profiling. Content-based filtering technique is to match properties of items of potential interest with stored user properties. Therefore, the more accurate and reliable the interest in the user profile, the more accurate the recommender system results. Jothis et al. (2019) claimed that the content-based filtering approach is an entity-specific algorithm that emphasises an entity's fundamental properties above its interactions with other entities to provide suggestions to the user. The flow in a content-based filtering recommender system is shown in Fig. 2. When users engage with the website, feedback is sent in order to learn about the user profiles. Before showing the user the top-ranked documents that are thought to be of the highest relevance, the recommender system will analyse the user profiles with the collection of the document.

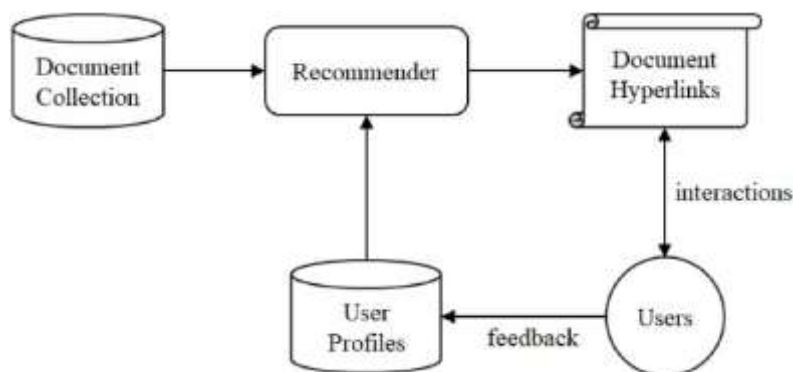


Fig. 2. Typical flow in Content-Based Filtering Technique (Kuo & Cheng, 2022)

Besides, a variety of models are used in the content-based filtering approach to determine similarity and produce appropriate suggestions. According to Isinkaye et al. (2015), utilizing Vector Space Model which is Term Frequency Inverse Document Frequency (TF-IDF), or a Probabilistic Model (Collaborative Filtering Technique), such as Naive Bayes, Cluster Analysis, Decision Tree, and Neural Networks, the

system assesses the relationship between various documents within a corpus. As a probabilistic model, the aforementioned machine learning methods are selected to forecast the likelihood of a user choice. As a result, under this strategy, the contents are given some priority. These ratings employ a variety of algorithms to determine the user's preferred and favourite content. Enough data and information are required to guarantee that this approach might be used effectively.

Content-based filtering technique uses object's metadata to generate accurate recommendation. Based on Jothis et al. (2019), this technique works best for contents such as research papers and articles. Object's metadata that are needed in a research paper, article or book are title, author, number of pages, publish year and other relevant information. By having sufficient information about the research paper, article or book, the system could recommend contents that the user might be interested in.

The content-based filtering approach has several benefits and drawbacks. The advantages of this technique are that it can foresee the proposal again in a reasonably short period of time if changes are made to the user profile. This method allows for creating suggestions for the user without the disclosure of user profiles, which may also safeguard the user's privacy. According to Al-Ghuribi & Noah (2019), explaining why a specific item is suggested might be simpler and offering the justification for the recommendations if a list of content features is used. In addition, content-based filtering approaches frequently offer highly personalised recommendations. This method also enables consumers to stop viral marketing and offers strong protection against the development of fraudulent items.

In contrast, content-based filtering techniques also have some disadvantages too. The biggest flaw with this method is the demand for in-depth comprehension and justification of the attributes of the information in the user profile. This technique requires complete and enough information in the user profile and sufficient information of the contents due to this technique highly relies on the object's metadata. Therefore, the amount of information will affect the efficiency of the filtering technique. Isinkaye et al. (2015) stated that content overspecialization as another significant issue with this approach. Users can only receive suggestions for things that match those they have previously indicated in their profiles. Due to the user profile constraint on the description of comparable things, the over-specialization issue results in users notobtaining new or different sorts of items.

2.2.2. Collaborative Filtering Technique

One of the techniques utilised in the recommender system is the collaborative filtering technique. Collaborative filtering techniques try to obtain a group of users based on their past actions or activities between users and items then make the suitable recommendation. This technique can be done by identifying users and items that have similar interests or behaviours. An illustration of a collaborative filtering technique is one that makes recommendations according to the interests of other people who are having the same interests. This technique functions by creating a user-item matrix containing user preferences for things (Isinkaye et al., 2015). It will then calculate the similarities between the users' profiles and combine users with similar interests in a neighbourhood group. Therefore, users could get recommendations to contents that he/she does not interact with, but the contents have been rated positively by other users in the same neighbourhood.

Memory-based collaborative filtering works with recorded user-item interactions that are stored in the matrix. A memory-based algorithm's key operations include calculating how similar users and items are, taking into consideration the user-item interactions matrix, and predicting the user rating. According to Jothis et al. (2019), user feedback as both explicit and implicit about an item is crucial for neighbour search. Suppose the user likes the neighbour's item and does not dislike things that are similar to the neighbour's preferences. In that case, the algorithm can use this neighbour to find the unusual region in a Venn diagram and offer a piece of the non-overlapping part. Al-Ghuribi & Noah (2019) asserted that the memory-based collaborative filtering method is an algorithm that heuristically forecasts the item's rating according to user ratings.

A user will be given suggestions of products loved by users who are mostly equivalent to them in collaborative filtering method with user-based and the goal is to seek out a neighbourhood of users who are similar and determine how similar they are by contrasting their prior interest or activity on the same things then provide recommendations to the user.

By comparing the rating of the same user gave each item, the collaborative filtering technique with item-based will determines how similar two identical items are. If most people that engaged with two items did so in a similar way, then the two items are said to be comparable. Users will receive recommendations of the similar items that they have been interested in the past. For example, suppose user provides a product a positive rating. In that case, the system will then suggest some products that are comparable to the product that has previously received a positive rating from the user.

Model-based collaborative filtering algorithms presumptively construct user and item representations based on models. By analysing user-item matrix, model-based filtering approaches frequently understand the relationship between users and items. Moreover, using ratings of previously rated products, the probabilistic technique is used to estimate the likelihood that a user would give a new item a particular rating. These probabilistic algorithms include association rule, clustering, decision tree, neural network, link analysis, regression and others are used to forecast the user's rating of some unseen items.

There are some advantages of using collaborative filtering techniques. According to Isinkaye et al. (2015), one of its benefits is that it may provide serendipitous and unfamiliar recommendations or suggestions for related items to the user even if the information is not contained in the profile. Additionally, collaborative filtering is adaptable enough to function in a range of domains, making it useful in those where there is minimal data associated with things.

Collaborative filtering techniques have some disadvantages too. The disadvantages include data sparsity, cold start, scalability and synonymy.

Data Sparsity Problem: When there is insufficient data to allow the system to function, data sparsity issues arise. The dataset's information on the person, the item, or both was lacking or incomplete when filling up the user-item matrix. This could reduce the system's performance to give suggestions and might increase the likelihood that it will make mistakes in its comments. Problems with data sparsity invariably result in a sparse user-item matrix and the system's inability to find pertinent neighbors.

Cold Start Problem: When neither the user nor the item has any information, a cold start problem occurs, which prevents the system from generating relevant recommendations. The user-item matrix's cells will have null values as a result of this issue. Consequently, the algorithm could not provide trustworthy suggestions for brand-new users or products.

Scalability: Rapid growth in the quantity of users and products in a system might cause scalability issues. As the dataset volume increases, a recommendation approach that is successful with a small number of datasets may not be able to continue to be effective. According to (Isinkaye et al., 2015), methods used to overcome scalability challenges and improve the development of proposals are built on reducing the dimensionality that is the Singular Value Decomposition (SVD) technique. This method is capable of producing effective and trustworthy recommendations.

Synonymy: The practise of assigning items with a great deal of similarity to various names or entries is known as synonymy. According to Isinkaye et al. (2015), several methods can be used to solve the synonymy problem, such as automated word expansion, development of thesaurus, and Singular Value Decomposition (SVD).

2.2.3. Hybrid Filtering Technique

A hybrid filtering strategy combines various ways of recommendation to create a more effective

solution. The most common hybrid filtering strategies are content-based filtering and collaborative filtering. Hybrid filtering technique can have better performance because it can overcome the shortcomings of each recommender method, hence forming a better recommender system. According to Isinkaye et al. (2015), combining several approaches by executing the algorithms separately and combining the results is possible, or by utilising collaborative filtering in a content-based approach and vice versa. Developing a uniform recommendation system combining both approaches is also one way to produce a hybrid filtering strategy. There are several hybridization methods such as weighted, mixed, cascade, switching, feature-combination, feature-augmentation and meta-level (Isinkaye et al., 2015).

Weighted hybridization creates a recommendation list or prediction by combining the results of different recommenders and combining the scores for each technique. Mixed hybridization combines recommendation results from different recommendation methods instead of having them at the same time. The cascade hybridization techniques apply an iterative refinement process to build an ordering of preferences between different items. The recommendations of one method are refined by another recommended method. In the switch hybridization, it has the ability to avoid problems specific to one method e.g. the new user problem of content-based recommender, by switching to a collaborative recommendation system. In the feature-combination, the features produced by a specific recommendation technique are fed into another recommendation technique. In the feature-augmentation, the technique makes use of the ratings and other information produced by the previous recommender. Finally, in the meta-level, the internal model generated by one recommendation technique is used as input for another.

2.3. Related Works

Al Alshaikh et al. (2017) described a unique collaborative filtering method that computes user similarity based on user profiles expressed as Dynamic Normalised Tree of Concepts (DNCTC) models rather than user ratings. Furthermore, recommendations are made using a community-centric tree (CCT) of concepts and a group of papers that might be relevant to the user's potential future interests are suggested using the CCT. BibSonomy dataset is a dataset that includes accurate records of users' interests as posts for academic publications throughout the course of about ten years. Each post includes the date, time, and information for a research paper. This study mainly used user data for computer users for years, 2015 and 2016, which included 1,642 individuals and 43,140 research publications. For the top N recommended papers, the precision for cut-off findings at position N (PN) is utilised as a criterion. One of the evaluation measures also takes into account the mean average precision across all users. With a MAP of 0.25, it is clear that the DNCTC model has the lowest precision performance.

A hybrid approach to the recommendation of scientific papers was developed by Amami et al. (2017). It blends concepts from collaborative filtering based on a relevance-based language model with content analysis based on probabilistic topic modelling. In order to bring like-minded academics together, a community detection technique will be used to build the researcher profiles based on the subjects retrieved by Latent Dirichlet Allocation (LDA) from the publications they have assessed. For this investigation, the 1.5 million DBLP4 publications and 700,000 researchers' dataset from ArnetMiner3 were employed. Only publications with complete titles and abstracts will be chosen from the dataset during preprocessing. According to early findings, the suggested method for recommending scientific papers, which combines topic models and relevance modelling, provides higher average Recall@m values than the RM + k-NN model and the PageRank-weighted CF model.

The collaborative strategy for the research article recommender system is presented in (Haruna et al., 2017). In order to tailor recommendations by utilising the pros of collaborative filtering technique, context metadata that is publicly available was used to discover the underlying relationships between papers. The proposed recommender system provides personalised suggestions regardless of the research field or the user's expertise. This study used data that included the publications of 50 researchers with interests in information retrieval, software engineering, user interface, security, graphics, databases, operating systems, embedded systems, and programming languages, among 1592

other areas of computer science. The evaluation metrics will be used are precision, recall, F1 measures, mean average precision (MAP), and mean reciprocal rank (MRR). For all values of N, the suggested method has consistently surpassed the baseline methods in evaluation metrics of precision. Additionally, based on recall and F1-score performance measurements, context-based collaborative framework outperforms the indicated technique in a list of 5 recommendations. However, the stringent criteria for selecting a candidate paper are the main cause of the suggested approach's poor recall performance. The baseline approaches based on MAP and MRR were significantly outperformed by the proposed method in all circumstances by suggesting the pertinent recommendations at the top.

Wang et al. (2018) presented a computer science-focused journal and conference recommender system. Real-time online system is built using softmax regression and chi-square in conjunction with content-based filtering. An automatic web crawler has been developed in order to collect the abstracts and other important information. The China Computer Federation (CCF) ranked 28 journals and 38 conferences as A-class, and the "root links" of their home pages were collected and used in this research paper. There are two different sorts of suggestion results offered by the suggested system which are one-class (Top 1) and three-class (Top 3). A number of feature choices, including chi-square, mutual information, and information gain, were assessed. The accuracy, F-measure, and ROC evaluation formula was used to assess the recommender system. The evaluation for feature selection in three classes yields better results than in one class. In this research, the chi-square model had the highest accuracy and F-measures. For instance, the Top3 target may surpass the Top1 goal by 75.2%, with the accuracy of 61.37% in chi-square-based classification.

A content-based method for suggesting citations in an academic article is presented by Bhagavatula et al. (2018) which allows researchers to conduct efficient literature research even when there is lack of data. The employed technique uses a neural model to encode the text of each available document, then chooses the closest neighbours of a query document as candidates then reranking the candidates using a second model. DBLP and PubMed datasets were used in this research with an average of 5 citations per article, the DBLP dataset contains more than 50K research articles in the field of computer science. Over 45K scholarly papers in the medical fields make up the PubMed dataset, and each item receives an average of 17 citations. Article's title, abstract, venue, authors, citations, and key phrases are all included in both datasets. Mean Reciprocal Rank (MRR) and F1@20 are employed as evaluation measures. With relative improvements of over 18% in F1@20 and over 22% in MRR, the proposed method exceeds the best results on the PubMed and DBLP datasets.

Haruna et al. (2020) extended their work published in Haruna et al. (2017) by proposing an adaptation based on the hidden associations that exist between research papers. Unlike existing approaches, they suggested an independent research paper recommendation framework that does not require a priori user profiles. The framework consists of the extractor, which uses the pre-filtering method to extract the contextual metadata including the references, citations, abstract and the title of the target paper. Then, the extracted information is analyzed through multiple calculations in the synthesis stage and finally the top-N recommendations are presented to the user. In their approach, they employed the publicly available dataset from DBLP. To prove the viability of their proposed framework, they conducted the experimental evaluation based on Precision, Recall and F1 against the content-based and collaborative approaches.

Zhu et al. (2021) developed a research paper recommender system based on the information retrieval approach. They represented the textual data using several approaches ranging from traditional term-frequency based methods and topic-modeling to embeddings. Then, the relevant information were transformed into vector representation. They employed the publicly available dataset from PubMed. In the evaluation, several evaluation metrics were computed such as Mean reciprocal rank (MRR), Recall, Precision, Mean average and precision (MAP).

Sterling and Montemore (2021) created a research-paper recommender system known as ExCiteSearch that uses citations and abstract similarity. It finds closely related papers to a set of related papers. The effectiveness of ExCiteSearch's search method is demonstrated by its ability to 1593

reproduce many of the reference lists-based similarity metric for scientific articles by clustering unsupervised sets of articles. They extracted the data from the Google Scholar as the dataset. The evaluation is conducted via similarity measure and represented in the heatmaps.

Zhang and Zhu (2022) studied the citation recommendation by focusing on the perspective of citing paper to cited paper. In another words, they based on the co-citation relationships among cited papers to denote cited paper's relations. Their methodology can be summarized as follows. First, extraction of co-citation relationships and citation content; second, representation of the citation content and citation relationships; third, calculation of similarity among the cited papers; and fourth, quantitative evaluation on the proposed approach. They employed the PLOS ONE dataset to extract the co-citation relationship and citation content through data parsing. The parsed The MySQL database is employed as the storage and processing medium. A total of 115,653 citation relationships are obtained, in which part of citation relationships may be repeated because a reference may be cited multiple time by the same citing paper.

For each of the three recommender system methodologies, numerous pertinent works have been assessed. For evaluation metrics, the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Precision, Recall, and F1 measurements were the most often used techniques in the evaluation part. In terms of MAE value, the model is more accurate when the MAE value is near to 0. On the other hand, for RMSE values between 0.2 and 0.5 will indicate that the proposed model can reasonably forecast the data reliably. Besides, the greater the value, the more accurate the model is for the Precision, Recall, and F1 score. Table 1 depicts the summary of the related works.

Table 1. Summary of Related Works

Article	Description	Dataset	Evaluation Metrics
(Al Alshaikh et al., 2017)	By assessing the similarity between users based on the users' profiles expressed as Dynamic Normalised Tree of Concepts (DNTC) models, this study introduced a novel collaborative filtering method that is independent of user ratings.	Dataset: BibSonomy (2015 and 2016) This dataset provides actual recordings of users' interests as postings for research papers.	The top N recommended papers are evaluated using the precision. One of the evaluation measures also takes into account the mean average precision across all users. With a MAP of 0.25, it is clear that the DNTC model has the lowest precision performance.
(Amami et al., 2017)	Present a hybrid method for recommending scientific papers that combines content analysis based on probabilistic topic modelling with concepts from collaborative filtering based on a relevance-based language model.	Dataset: ArnetMiner3's dataset, which includes 1.5 million DBLP4 papers and 700,000 researchers.	According to the preliminary findings, the suggested method, which integrates topic models and relevance modelling into the process of recommending scientific papers, produces superior average Recall@m values than the RM + k-NN model and the PageRank- weighted CF model.
Haruna et al., 2017)	This paper presents the consultative technique for the research article recommender system. The suggested recommender system offers tailored recommendations independent of the user's level of experience or the study field.	Dataset: DBLP This study used a dataset that included the publications of 50 researchers with interests in a few areas such as software engineering, databases and etc.	Precision, recall, F1 measures, mean average precision (MAP), and mean reciprocal rank (MRR) are used to assess the recommender system. The model is examining by using precision, recall, F1 measures, mean average precision (MAP), and mean reciprocal rank (MRR). The proposed method greatly outperforms the baseline approaches based on MAP and MRR in all.

(Wang et al., 2018)	Proposed a computer science-focused journal and conference recommender system. A real-time online system is built using chi-square and softmax regression in conjunction with content-based filtering.	Dataset: Gathered through web crawler on the abstracts and other pertinent data. The China Computer Federation (CCF) ranked 28 journals and 38 conferences as A- class, and the "root links" of their home pages were collected and used in this research paper.	Two different sorts of suggestion results are offered by the suggested system: one-class (Top 1) and three-class (Top 3). A number of feature choices, including chi-square, mutual information, and information gain, were assessed. The accuracy, F-measure, and ROC evaluation formula was used to assess the recommender system. The evaluation for feature selection in three classes yields better results than in one class.
(Bhagavatula et al., 2018)	This study proposes a content-based citation recommendation system that is flexible for researchers to conduct literature review even when querying documents lacking data.	Dataset: DBLP (computer science domain) and PubMed (medical domain)	The evaluation metrics used are Mean Reciprocal Rank (MRR) and F1@20. With relative improvements of over 18% in F1@20 and over 22% in MRR, the suggested strategy exceeds the best results on the PubMed and DBLP datasets without the use of metadata.
Haruna et al. (2020)	This study proposes an adaptation based on the hidden associations that exist between research papers by employing the pre-filtering method to extract the contextual metadata.	Dataset: DBLP	The evaluation metrics are based on Precision, Recall and F1.
Zhu et al. (2021)	This study proposes research paper recommendation based on the information retrieval approach. They represented the textual data using several approaches ranging from traditional term- frequency based methods and topic-modeling to embeddings.	Dataset: PubMed	In the evaluation, several evaluation metrics were computed such as Mean reciprocal rank (MRR), Recall, Precision, Mean average and precision (MAP).
Sterling and Montemore (2021)	This study creates a research-paper recommender system that uses citations and abstract similarity to find closely related papers to a set of related papers. In addition, they demonstrated its ability to reproduce many of the reference lists-based similarity metric.	Dataset: Extracted from Google Scholar.	The evaluation is conducted via similarity measure and represented in the heatmaps.
Zhang and Zhu (2022)	This study focuses on citation recommendation based on the co-citation relationships and citation content.	Dataset: Extracted from PLOS ONE under the domain of Artificial Intelligence in 2018.	Evaluation metric: AUC, MAP and a case study

Despite the growth and development of many recommender systems for e-commerce, movies, books, and other media, there is still a significant gap in the development of machine learning-based algorithms and high-performance systems to find the best suggestion. Currently there are little to no recommender systems that make recommendations on research articles as mostly will focus more on e-commerce or movies. According to (Pujahari & Sisodia, 2022), content-based recommendations are a common practice in contemporary recommendation contexts including music, movies, and other online resources. According to (Kuo & Cheng, 2022), content-based filtering is more likely to experience overspecialization as a result of recommendations with

similar qualities when the predicted items are strongly connected to the user's previous interactions with the goods, which limits the novelty. Therefore, hybrid recommender systems that mix content-based filtering with collaborative filtering approaches are typically employed to overcome each other's shortcomings in order to address this problem.

3. PROPOSED SYSTEM

The prototype will first input the required dataset, which are the article.csv, author_list.csv, author_write.csv and author_cite.csv. Data cleaning will be carried out for those datasets that need it once all the other datasets have been inserted. The next step will be training a recommender system with the cleaned dataset. Next, a prediction or suggestion will be displayed. Further, the cosine similarity of the predicted results will be shown as well. Fig. 3 shows the flowchart of the implemented prototype.

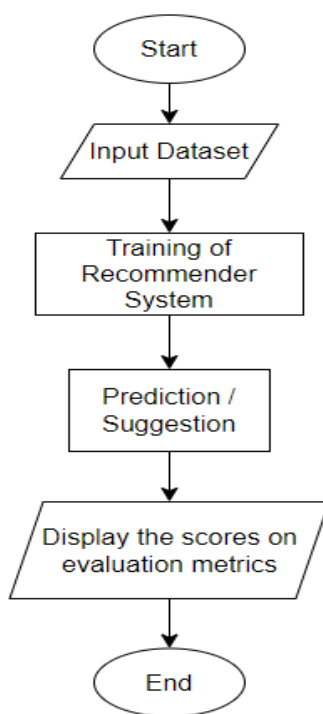


Fig. 3. Flowchart of Prototype

The suggested recommender system is a hybrid-based recommender system which implement mixed hybridization by merging the collaborative filtering technique with the content-based filtering technique. The content-based filtering technique will suggest articles based on the article's feature such as title and keywords. Collaborative filtering analyses the article's list of citations and offers suggestions based on comparable citations. The user will receive a wider range of recommendations after combining the recommendation results from the collaborative filtering technique and the content-based filtering technique.

Combining these methods can help users receive recommendations that are more precise and comprehensive while also overcoming the shortcomings of individual approaches. Hybrid-based recommender systems can be used to increase recommendation accuracy and provide users with more pertinent choices. Additionally, different recommendation techniques may perform better in various circumstances. Mixed hybridization can increase the coverage of the recommender system and offer recommendations for a wider range. It also aids in adding diversity to recommendations, giving users access to a wider range of suggestions. On top of that, adopting mixed hybridization makes recommender systems more reliable because other techniques can step in and make recommendations if one recommendation method performs poorly.

3.1. Dataset

The Scopus API will be utilised to obtain the dataset for this paper. Scopus is an Elsevier's abstract and citation database which collects articles published in almost all scientific journals. Developers can automatically retrieve data from Scopus using the API. The API allows developers to write programs that automatically extract data from Scopus and add that data to their system (see Fig. 4).

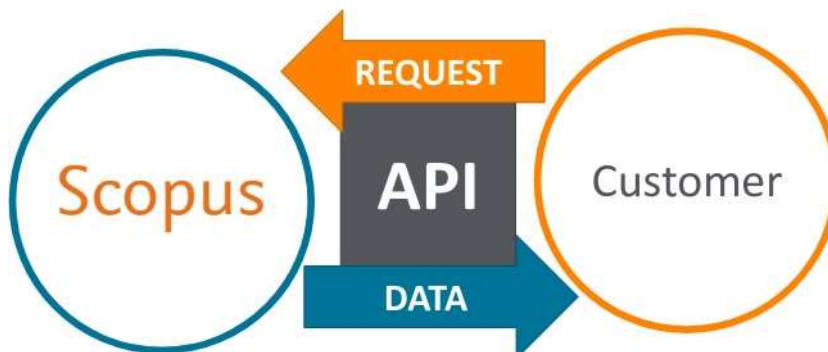


Fig. 4. Scopus API

The Scopus data model is designed around the notion that articles are written by authors that are affiliated with institutions. Using this data model, Scopus has provided a wide variety of API including the Search API, Retrieval API and Metadata API. The Search API allows users to select a list of results (scopus, affiliation or author) based on a certain query. The Retrieval API allows users to retrieve the information about a certain scopus/abstract, affiliation or author.

The dataset was divided into 8 independent CSV files.

- article.csv
 - Article's id, which is unique in whole dataset (article_id)
 - Article eid in scopous (eid)
 - Article Type (aggregationType)
 - Article Abstract (abstract)
 - Article Title (title)
 - Article DOI (doi)
 - Article Keywords, in array [], separated by "," (keywords)
- author_1.csv, author_2.csv, author_3.csv, author_4.csv, author_5.csv, author_6.csv
 - Author's id, which is unique in this dataset (author_id)
 - The last institution name record in scopus (latestAffiliatedInstitution_name)
 - author e-id in scopus (eid)

- author id in scopus (authorId)
- author first name (preferredName_first)
- author last name (preferredName_last)
- author full name (preferredName_full)
- author published document count (documentCount)
- co-author document count (coAuthorsCount)
- author's article citation count (cited by other) (citationsCount)
- author's published subject area, array (publishedSubjectAreas)
- author's email address (emailAddress)
- author name array, might contains more than one (nameVariants)
- author hindex (hindex)
- contains link url to retrieve information from scopus, like author scopus
- profile, published document, co-author document and etc (links)
- same as citationsCount, just the old record, can ignore (citedByCount)
- author's latest institution id in scopus (latestAffiliatedInstitution_id)
- author's latest institution city (latestAffiliatedInstitution_address_city)
- author's latest institution state (latestAffiliatedInstitution_address_state)
- author's latest institution country (latestAffiliatedInstitution_address_country)
- author's latest institution links url (latestAffiliatedInstitution_links)
- author_write.csv
 - Author's unique id in dataset (author_id)
 - Article's unique id in dataset (article_id)

3.2. Data Cleaning

Before implementing any algorithms, data cleaning is a crucial step since clean data will potentially boost overall efficiency and have higher quality information. Data cleaning is needed in all CSV files, including "article.csv," "author_1.csv," "author_2.csv," "author_3.csv," "author_4.csv," "author_5.csv," and "author_6.csv."

Initially, there is a single CSV file with the name "article.csv." This data frame has the shape (152045, 7), indicating that it has 152045 rows and 7 columns. Only 50,000 articles will be included in the dataset in order to increase the recommender system's effectiveness. All related files will likewise be reduced so that they just contain the data from the selection of 50,000 articles. Using a SQL query, the data will be chosen, truncated, and exported

into a new CSV file.

Following that, it is crucial to check the data set to determine whether there are any null values in the 'article.csv'. The outcome of the count of the null value in each column is displayed in Fig. 5 To ensure that the recommender system continues to function without issue, all null values will thereafter be handled by inserting an empty string.

article_id	0
eid	0
aggregationType	107
abstract	6326
title	30
doi	3605
keywords	0

Fig. 3. Null Count in article.csv

In addition, the 'author_1.csv', 'author_2.csv', 'author_3.csv', 'author_4.csv', 'author_5.csv', and 'author_6.csv' CSV files containing all the authors' information will be reviewed. Prior to truncating the data using SQL, the six CSV files related to the authors will be concatenated into a single file with the name "author_list.csv". Combining all 6 CSV files as one is preferred since all 6 CSV files include different data rows but not different columns. Next, checking on the null value will be carried out and Fig. 6 shows the results of the null values in each column. The null values will then be handled using the same procedure which is replacing them with an empty string.

A cleaned dataset on the author will then be filled with all the author data after processing each null value in the author data frame. The newly dataset will be utilised for further implementation after being successfully cleaned.

3.3. Recommender System

After cleaning all the necessary datasets, the data set is ready to be used in the algorithms that will perform the recommender system. The prototype includes both collaborative filtering and content-based filtering, a total of two distinct recommender systems. There are a few columns in the article dataset, and the recommender system will offer suggestions for content-based filtering based on the article's title and keywords. The collaborative filtering technique will use the article's list of citations to determine which articles are related to one another in terms of citations.

author_id	0
latestAffiliatedInstitution_name	7190
eid	0
authorId	0
preferredName_first	1126
preferredName_last	392
preferredName_full	392
documentCount	0
coAuthorsCount	0
citationsCount	0
publishedSubjectAreas	0
emailAddress	147203
nameVariants	0
hindex	0
links	0
citedByCount	0
latestAffiliatedInstitution_id	7190
latestAffiliatedInstitution_address_city	13883
latestAffiliatedInstitution_address_state	431101
latestAffiliatedInstitution_address_country	10502
latestAffiliatedInstitution_links	7190

Fig. 6. Null Count in author data frame

In order to use content-based filtering, the dataset with only the title and keywords column is fitted with the relevant TF-IDF model, creating an embedding vector with huge dimensions for each term in each article. The user-selected article's title and keywords will also be entered into a different TF-IDF model. Then, using cosine similarity, each one will be compared to the vectors representing the title, keywords, and the chosen article in the same vector space. The ranking function is carried out as the next step. It will choose the best index from the distance matrix and take the cosine similarity distance to arrange them from highest to lowest. The top 3 articles, together with the corresponding article ID, title, abstract, keywords, and cosine similarity, will then be shown. The results of the suggestion for title and keywords will be merged, and any duplicate suggestions will be removed. Additionally, recommendations that have a cosine similarity value of less than 0.5 will be removed because there will be an assumption of the results for the recommended article must have a minimum threshold of 0.5 for cosine similarity value.

For collaborative filtering technique, an array with the column and row size of the total of the article will be created. Then, those article that are being cited will be considered as related and will be assigned for a weight of 1. This will be applicable to all the article and will be used in the comparison with the selected article by the user. Another array will be created to contain the information of the citing list of articles in order to compare with the array that are having the list of citation for all the article. Then, using cosine similarity, each one will be compared to the vectors and the chosen article in the same vector space. The ranking function is carried out as the next step. It will choose the best index from the distance matrix and take the cosine similarity distance to arrange them from highest to lowest. The top 3 articles, together with the corresponding article ID, title, abstract, keywords, and cosine similarity, will then be shown. Additionally, recommendations that have a cosine similarity value of less than 0.5 will be removed because there will be an assumption of the results for the recommended article must have a minimum threshold of 0.5 for cosine similarity value.

3.4. User interface

The user interface is as shown in Fig. 7 once the program is run. User must key in their username and password in order to log in to the system. If the user does not have any account, then they must sign up for a new account by clicking the sign-up button.

After login, user will be redirected to the main menu of the system which have two option for the user to choose

which are “Search Article” and “Search Author” as show in Fig. 8.



Fig. 7. Login Page



Fig. 8. Main Menu Page

If the user clicks on the “Search Article”, it will then be redirected to the page as shown in Fig. 9 and user will just have to type in any keywords in order to search on the relevant article.



Fig. 9. Search Article Page

If the user clicks on the “Search Author”, it will then be redirected to the page as shown in Fig. 10, which the user must type in information in either the author’s first name or the author’s last name.



Fig. 10. Search Author Page

Once all the details have been inserted and user has selected the article, the article details will then be shown as in Fig. 11.



Fig. 11. Article Details Page

After that, user can click on the button “PROCEED to RECOMMENDED ARTICLE” to proceed with showing a list of recommended articles with the article ID, article title and cosine similarity for content-based filtering technique and collaborative filtering technique (see Fig. 12).



Fig. 12. Recommended Article Details Page

To continue, user can select one of the articles and click on the select article button. The system will then redirect to a page that containing the authors' email address of the selected article, the recommended article ID, title, DOI, keywords and abstract as shown in Fig. 13.



Fig. 13. Email Template Page

3.5. Evaluation and Preliminary Results

In this research, there are two different techniques have been implemented which are content-based filtering technique and collaborative filtering technique, hence the findings will include evaluation metrics on cosine similarity for both techniques.

First, the evaluation metrics for content-based filtering will be discussed. The cosine similarity score for the top 3 articles based on the content-based filtering technique is displayed in Table 2. "Smart city air quality prediction using machine learning" are the search terms for title and "Multi- Layer Perceptron, Air Pollution, Random Forest, Particulate Matter (PM2.5)" are the search terms for keywords. They should have a cosine similarity of 1 with the first suggested article because they share exactly the same phrases.

Table 2. Cosine Similarity Results on Top 3 Recommended Article for Content-Based Filtering Technique

	First	Second	Third
Cosine Similarity Value	1	0.65	0.53

The first article, which has the same keywords as the query sentence, is shown in Fig. 14 together with the other two recommended articles.

article_id	title	abstract	keywords	cosine_similarity
0 994	Smart city air quality prediction using machin...	© 2021 IEEE Air pollution in smart cities in t...	[Multi-Layer Perceptron, Air Pollution, Random F...	1.000000
1 118916	Time Series Analysis and Forecasting of Air Po...	© 2013 IEEE Current development of Pakistan's ...	[PM2.5, Particulate matter, air pollution, PM10]	0.653162
2 17320	Using machine learning to forecast air and wat...	© 2021 by SCITEPRESS - Science and Technology ...	[Deep learning, Machine learning, Environmental ...	0.532298

Fig. 14 Results of Recommended Article for Content-Based Filtering Technique

Cosine similarity will also be one of the evaluation metrics for the collaborative filtering technique. Table 3 shows the cosine similarity score for the top 3 articles using the collaborative filtering method. The first recommended article should have a cosine similarity of 1 because it shares the same list of citations as the selected article, which has the ID 13295.

Table 3. Cosine Similarity Results on Top 3 Recommended Article for Collaborative Filtering Technique

	First	Second	Third
Cosine Similarity Value	1	0.85	0.71

The first article, which shares the same list of citations as the chosen article, is displayed alongside the other two suggested articles in Fig. 15.

article_id	title	abstract	keywords	cosine_similarity
1746	13295 The relationship between renewable energy and ...	© 2021 by the authors. Licensee MDPI, Basel, S...	[Covid-19,Brazil,Artificial neural networks,Ec...	1.000000
1747	13300 Using an artificial neural networks experiment...	© 2021 by the authors. Licensee MDPI, Basel, S...	[Artificial Neural Networks,TFP,Agricultural s...	0.857143
1753	13339 The nexus between COVID-19 deaths, air pollut...	© 2021 The Author(s)The aim of this paper is 1...	[COVID-19,Air pollution,New York state,Machine...	0.714286

Fig. 15 Results of Recommended Article for Collaborative Filtering Technique

The list of citations, which includes a total of 7 articles, is presented in Fig. 16 and the selected article's ID is 13295. The third suggested article ID is 13339, and Fig. 17 displays the list of citations, which includes a total of 7 articles. Due to the fact that the third suggested article has 5 overlapped articles from the chosen article, its cosine similarity value is 0.71.

article_id	cited_article_id
13295	92407
13295	92405
13295	77778
13295	77774
13295	77773
13295	66695
13295	13312

Fig. 16. Citation List for Article 13295

article_id	cited_article_id
13339	104104
13339	92407
13339	77778
13339	77773
13339	74498
13339	66695
13339	13312

Fig. 17. Citation List for Article 13339

CONCLUSION AND FUTURE WORK

Several recommendation methods, including content-based filtering, collaborative filtering, and hybrid filtering, have been studied in this research work. In addition, a review of the literature highlighting current research on a range of recommender systems using various recommender filtering methods is being explored and discussed. A table summary of related works is also included, along with the description, dataset and evaluation metrics of the related works on the recommender systems.

In order to further develop the current recommender system, we need carry out more evaluations in the future. These evaluations will allow us to compare the performance of the recommender system to other that are more recent works.

REFERENCES

- [1] Zhou, C., Leng, M., Liu, Z., Cui, X., Yu, J. (2022). The impact of Recommender Systems and pricing strategies on Brand Competition and consumer search, *Electronic Commerce Research and Applications*, vol. 53, 101144.
- [2] Zaveri, A.A., Mashood, R., Shehmir, S., Parveen, M., Sami, N., Nazar, M. (2023). AIRA: An Intelligent Recommendation Agent Application for Movies, *Journal of Informatics and Web Engineering*, vol. 2(2), pp. 72-89, 2023.
- [3] Elahi, M., Kholgh, D. K., Kiarostami, M. S., Saghari, S., Rad, S. P., Tkalcic, M. (2021). Investigating the impact of recommender systems on user-based and item-based popularity bias. *Information Processing & Management*, vol. 58 (5), 102655.
- [4] Chew, L. J., Haw, S. C., Subramaniam, S. (2020). Recommender System for Retail Domain: An Insight on Techniques and Evaluations. *International Conference on Computer Modeling and Simulation*, 9-13.
- [5] Karimi, M., Jannach, D., Jugovac, M. (2018). News Recommender Systems – survey and roads ahead. *Information Processing & Management*, vol. 54(6), 1203-1227.
- [6] Khademizadeh, S., Nematollahi, Z., Danesh, F. (2022). Analysis of Book Circulation Data and a book recommendation system in academic libraries using data mining techniques. *Library & Information Science Research*, vol. 44(4), 101191.
- [7] Melville, P., Sindhvani, V. (2017). Recommender Systems. In: Sammut, C., Webb, G.I. (eds) *Encyclopedia of Machine Learning and Data Mining*. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7687-1_964
- [8] Joe, C. V., Raj, J. S. (2021). Location-based orientation context dependent recommender system for users. *Journal of Trends in Computer Science and Smart Technology*, vol. 3(1), 14-23.
- [9] Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, vol. 16(3), 261–273.
- [10] Falk, K. (2019). *Practical recommender systems*. Manning Publications Company, O'Reilly.
- [11] Ziauddin, I., Khan, M., Jam, F., & Hijazi, S. (2010). The impacts of employees' job stress on organizational commitment. *European Journal of Social Sciences*, 13(4), 617-622.
- [12] Gulzar, Z., Leema, A. A., & Deepak, G. (2018). PCRS: Personalized course recommender system based on Hybrid Approach. *Procedia Computer Science*, vol. 125, 518-524.
- [13] Sharma, R., Gopalani, D., & Meena, Y. (2023). An anatomization of research paper recommender system: Overview, approaches and challenges. *Engineering Applications of Artificial Intelligence*, vol. 118, 105641.
- [14] Jothis, K. R., LokeshKumar, R., Anto, S., Gauri T. (2019). Qualitative analysis of models and issues in Recommender Systems. *Journal of Computational and Theoretical Nanoscience*, vol.16(5-6), 1881-1888.
- [15] Kuo, R. J., Cheng, H.R. (2022). A content-based recommender system with consideration of repeat purchase behavior. *Applied Soft Computing*, vol. 127, 109361.
- [16] Al-Ghuribi, S. M., Noah, S.A.M. (2019). Multi-criteria review-based recommender system—the state of the art. *IEEE Access*, vol. 7, pp. 169446-169468.
- [17] Al Alshaikh, M., Uchyigit, G., Evans, R. (2017). Predicting Future Interests in a Research Paper Recommender System using a Community Centric Tree of Concepts Model. *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 91- 101.
- [18] Amami, M., Faiz, R., Stella, F., Pasi, G. (2017). A graph based approach to scientific paper recommendation. *International Conference on Web Intelligence*, 777–782.
- [19] Haruna, K., Ismail, M. A., Damiasih, D., Sutopo, J., & Herawan, T. (2017). A collaborative approach for Research Paper Recommender System. *PLoS ONE*, vol. 12(10), e0184516.
- [20] Wang, D., Liang, Y., Xu, D., Feng, X., & Guan, R. (2018). A content-based recommender system for Computer Science Publications. *Knowledge-Based Systems*, vol. 157, 1-9.
- [21] Pujahari, A., & Sisodia, D. S. (2022). Item feature refinement using matrix factorization and boosted learning based user profile generation for content-based Recommender Systems. *Expert Systems with Applications*, vol. 206, 117849.
- [22] Chávez, J. A. M.-., Arroyo, N. S. M. ., María, S. A. F. de ., Saravia, R. C. ., Hinostrroza, M. R. D. ., Montenegro, L. P. B. ., Toledo, M. F. M. ., Olivares, J. A. C. ., & Puga, N. B. . (2023). Artificial Intelligence in Engineering and Computer Science Learning: Systematic Review Article. *International Journal of Membrane Science and Technology*, 10(3), 221-233. <https://doi.org/10.15379/ijmst.v10i3.1514>
- [23] Kuo, R. J., & Cheng, H.-R. (2022). A content-based recommender system with consideration of repeat purchase behavior. *Applied Soft Computing*, vol. 127, 109361.
- [24] Philip, S., Shola, P., & Ovyte, A. (2014). Application of Content-Based Approach in Research Paper Recommendation System for a Digital Library. *International Journal of Advanced Computer Science and Applications*, vol. 5(10), <https://doi.org/10.14569/ijacsa.2014.051006>
- [25] Al Alshaikh, M., Uchyigit, G., & Evans, R. (2017). Predicting Future Interests in a Research Paper Recommender System using a Community Centric Tree of Concepts Model. *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, vol. OIC3K, 91-101.
- [26] Amami, M., Faiz, R., Stella, F., & Pasi, G. (2017). A graph based approach to scientific paper recommendation. *International Conference on Web Intelligence*, 777–782.
- [27] Haruna, K., Ismail, M. A., Damiasih, D., Sutopo, J., Herawan, T. (2017). A collaborative approach for Research Paper Recommender System. *PLOS ONE*, <https://doi.org/10.1371/journal.pone.0184516>

- [28] Wang, D., Liang, Y., Xu, D., Feng, X., & Guan, R. (2018). A content-based recommender system for Computer Science Publications. *Knowledge-Based Systems*, vol. 157, 1-9.
- [29] Bhagavatula, C., Feldman, S., Power, R., & Ammar, W. (2018). Content-Based Citation Recommendation. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 238-251.
- [30] Haruna, K., Ismail, M. A., Qazi, A., Kakudi, H. A., Hassan, M., Muaz, S. A., & Chiroma, H. (2020). Research paper recommender system based on public contextual metadata. *Scientometrics*, 125, 101- 114.
- [31] Zhu, J., Patra, B. G., & Yaseen, A. (2021). Recommender system of scholarly papers using public datasets. *AMIA Summits on Translational Science Proceedings*, 672.
- [32] Sterling, J. A., & Montemore, M. M. (2021). Combining citation network information and text similarity for research article recommender systems. *IEEE Access*, vol. 10, 16-23.
- [33] Zhang, J., Zhu, L. (2022). Citation recommendation using semantic representation of cited papers' relations and content. *Expert Systems with Applications*, vol. 187, 115826.
- [34] Pujahari, A., Sisodia, D. S. (2022). Item feature refinement using matrix factorization and boosted learning based user profile generation for content-based Recommender Systems. *Expert Systems with Applications*, vol. 206, 117849.

DOI: <https://doi.org/10.15379/ijmst.v10i2.1830>

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.