Clustering Based Recommender System for Compilation of Research Papers

Abstract: Due to addition of the journals and conference proceedings around the globe in the literature, it is observed that lots ofresearch papers are daily published in the various categories of the journals and conference proceedings. Researchers sometimes encounter difficulties while endeavoring to comprehensively analyze all research publications for a particular field of study to locate pertinent studies which may be used forfuture scope of the work over the research topic.Hence, in the present work, a recommender system is explained to the research scholars by proposing articles based on assessments supplied by other academics within the similar domain of research. Collaborative filtering technique is used for the development of the recommender system, and it may be extensively utilized in several commercial recommender systems. It is obvious that the computational complexity of methods grows and directly proportion to the number of users and items. To address the said issues related to scalability, a proficient recommender system is presented that makes use of subspace clustering. The approach entails examining the researcher-paper matrix to ascertain the correlations among different researchers. Through the relationships among the keywords of the research papers, the present work offers a well selected compilation of research articles for recommendation which may be used for future research work.

Keywords: Research Papers, Collaborative Filtering, Recommender System, Subspace Clustering, Hash Table,Model-based Systems.

1. Introduction

From the literature, it is seen that the old research articles are publishing either in scan form or in electronic form and many of the journals and conferences are publishing the research articles either in online or offline for the use of the researchers. The suggestion for recommender system is a challenging task. But it became very easier due to available techniques of data mining and hence e-commerce and m-commerce are gaining popularity in the current days. For completion of the research, a course work is must and during the course work, researchers must collect the related articles pertaining to assigned research work. Although Google Scholar is providing vast facility of searching through keywords for accessing the relevant research articles, but this list contains thousands of research papers for one keyword, but a researcher must scrutinize all pertinent publications within assigned specific field of study. Subsequently, researchers need to ascertain and select the relevant research articles that are directly relevant to field of study activity. The researchers might take longer time for incremental collection of the research articles and hence will face the difficulty due to duplication of the research articles available online. There is need of selection of unique research articles pertaining to field of study. Therefore, recommender systems are developed by the researchers through the concepts of the data mining.In this regard, a recommender system [1] is a specialized tool for filtering the information that is designed to discover a selection of articles that will be of interest to a certain researcher. It compiles a selection of scholarly articles for a particular researcher, based on the preferences of other academics who have similar interests in the same subject. Recommender systems are primarily divided into Content Based Filtering (CBF) and Collaborative Filtering (CF) [2]. The proposal in CBF considers the intrinsic characteristics of the object. Profile of the user is created by including specific keywords or attributes. The scoring of items is determined via degree of conformity to the user's attribute profile, and the most suitable matches are recommended. The collaborative filtering approach employed a database containing user-assigned ratings or preferences for different items to predict new items or products that a user may find attractive. In datasets with many dimensions on the research articles, a certain proportion of dimensions which are not relevant, are not considered for the field of study. Subspace clustering algorithms [3] restrict the investigation of important dimensions, allowing for the detection of clusters that exist in several subspaces, perhaps overlapping with one another.

The work reported here presents a recommender system which helps to the researchers for gathering the research articles according to specified field of research. A concept of CF is implemented over the several research articles for the development of the recommender system which may be further refined according to sub keywords matching as per field of study. The complexity of the system is measured and presented in the form of tables and graphs. The relationship is established among the keywords and sub keywords and various parameters are observed for judging the efficiency of the proposed system.

2. Related Work

Let us describe some of the important research papers relate to the present article. Computational citation analysis enables scientists to assess the influence of a specific item inside the network of citations [4]. Referencing another article is a method employed to recognize the sources that have

influenced the study endeavor. Nevertheless, this impact is not readily and immediately evident throughout the citation process. Authors demonstrated appreciation for the contributions of others by referencing thepublications. Garfield's research study [5] examines the problems that arise from the lack of uniformity in citation standards, specifically on the inclusion of citation details that impact the quality of citations As stated by the authors [6], this discrepancy might result in inaccuracies throughout the analytical procedure while trying to determine if the journal was cited or not. MacRoberts and MacRoberts [7] examined the bibliographic challenges that might possibly be resolved using citation analysis. Both studies examine the variability of several factors, including citation formats, changes in citation rates, types of publications, knowledge transfer, and forms of specialization. Consequently, the majority of scholars focus their endeavors on citation analysis, which may be classified into three main categories: co-citation analysis, content analysis, and impact analysis. Academic scholars frequently utilize the content-based filtering approach [8]. The researcher mostly employed articles, authors, and locations that had established associations with them. The majority of research papers often utilize conventional terms as a standard for comparison, with 70-83% of the studies employing TF-IDF as the weighting system [9-10]. Furthermore, the Content-based filtering technique incorporates several other attributes associated with the objects, such as the n-grams linguistics model, layout information, social tag model, and themes relatedness. The majority of research have utilized the Vector Space Model to store information that defines the representations of objects and user models [11].

3. Collaborative Filtering Technique

The CF approach is used for ratings or votes provided by other users to approximate the usefulness of items for a particular user known as active user and is based on the premise that consumers who have demonstrated similar preferences in the past are likely to demonstrate similar preferences in future also. Recommender systems are utilizing CF technique [12] to rely on a database of user preferences to predict futurer items or products that may be of interest to a new user. This forecast is generated by considering the expressed preferences of other users for specified goods. The user-based CF approach seeks to anticipate the preferences of a given user by locating other users who exhibit similar rating patterns and leveraging the ratings on other items to produce predictions, as seen in the figure 1.

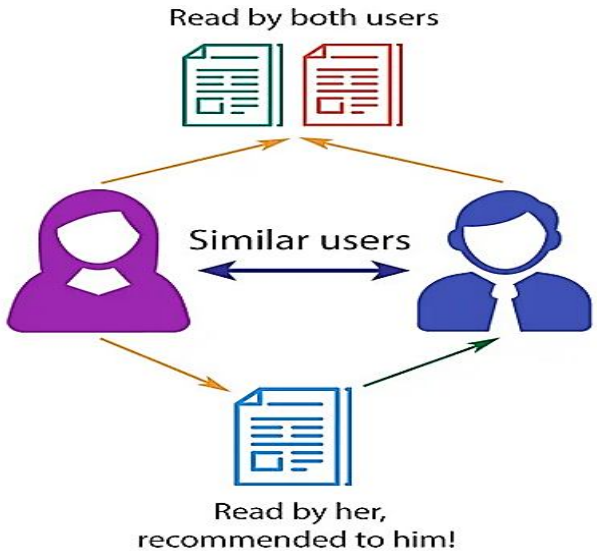

Figure 1. The Collaborative Filtering Technique

The item-based on CF approach, as explained in the reference, leverages the resemblances in rating patterns of items to generate predictions about user preferences. Memory-based CF methods utilize the whole database throughout each proposal generation phase. Model-based CF algorithms extract pertinent data from the dataset and employ it as a "model" to create recommendations without the necessity of processing the entire dataset for each proposal [13]. Nevertheless, the process of building a model sometimes necessitates a substantial investment of both time and resources. The integration of additional data into model-based systems frequently presents difficulties, leading to a deficiency in flexibility. In model-based collaborative filtering, there is no use of the whole dataset, which might lead to less precise predictions compared to model-based systems.

4. Subspace Clustering

Subspace clustering is an enhanced method that builds on traditional clustering methods by detecting clusters in many subspaces within a dataset [14]. One point may be a member of many clusters, each of which resides in a separate subspace. Traditional clustering methods examine all the dimensions of an input dataset which may be used for obtaining the comprehensive information. In real-world situations, certain dimensions may be irrelevant and might mask clusters within noisy data. Feature selection is a process that gets rid of unwanted and redundant dimensions by analyzing the entire dataset as presented in the figure 2. In the realm of research papers, the volume of scholarly articles exceeds the count of researchers.

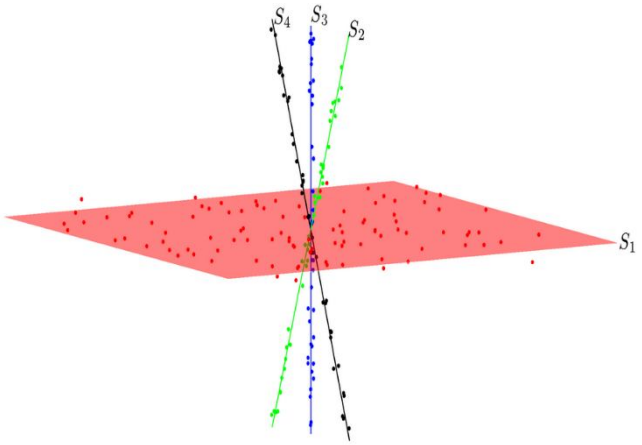

Figure 2. The Subspace Clustering

In certain fields, such as academic research, high-dimensional data often includes many irrelevant dimensions. Explicitly, the present work is not covering the research papers that have been assessed as below standard. Subspace clustering algorithms employ only relevant dimensions to detect clusters that exist inside overlapping subspaces. There are many, perhaps subspace algorithms which may be categorized into two primary groups: top-down and bottom-up, based on the search methodology. Top-down techniques, such as PROCLUS and FINDIT, systematically examine the whole dimensional space and progressively search for ever smaller subspaces.Projected Clustering (PROCLUS) is a technique for dimension-reduction subspace clustering technique. It begins by finding an original approximation of the clusters in the high dimensional attribute area relatively starting from individual dimensional spaces.

At first, they do an exhaustive search for all dimensions included in the dataset. Nevertheless, within the framework of a research paper[15], thoroughly scrutinizing each dimension, particularly those that are insignificant, can be a time-consuming endeavor. For this, recommender system is not suited. Bottom-up techniques, such as Clustering in Quest (CLIQUE) and Maximal Frequent Itemset Algorithm (MAFIA), initially prioritize the identification of significant areas in a particular dimension. Higher dimensional subspaces are only explored when there is a potential for clusters to exist inside them.

5. Suggested Approach

One can overseethe actions of researchers to compile a database that encompasses details such as the researcher's identity, generation of research papers, and the rating given to the study by the researcher. A rating from 0-5 is used in the recommender system. A subspace cluster refers to a collective of researchers who has common interests in a certain field of study. The proposed solution adheres to a five-step methodology. When utilizing the system, the researcher must first enter the domain of theresearch. Afterwards, the collection of articles within that field is acquired and given to the user. The system is provided with a matrix [16] that depicts the association between researchers and publications. This matrix is specially tailored to the user's topic of study and contains ratings assigned by the researcher to the publications. The output comprises a compilation of highly rated research publications in the field, omitting that have been reviewed by the researcher in question.

The number of research publications published on the web surpasses the number of researchers. Here, there is need to eliminate dimensions that are not relevant. The presented work only evaluates that have gotten favorable evaluations. To tackle the problem of limited data in the field of research papers, the rows in the matrix representing the relationship between researchers and papers are transformed into strings that show the positions of the highly rated publications. For this scenario, ratings of 4 and 5 are classified as high ratings. For example, if a researcher awarded a rating of 2 to paper1, 4 to paper2, 5 to paper3, and 2 to paper4, then the appropriate row for the researcher is transformed into 2 and 3. The result of this translation is a set of strings that represent the id's of research articles that have been highly rated by researchers. Denote the changed dataset as T, where each row is identifiable by the row^{id}.

For each row in the transformed dataset T, do a pairwise comparison with all the rows that come after it. Instantiate a hash table and initializeall of its values to null. Should there be any intersection among rows, proceed to modify the intersecting area in a hash table, including its matching row identifier. If the intersection is already present in the hash table, then modify the count value accordingly. The converted data for researcher1 consists of the values 2, 4, 5, and 8, whereas the changed data for researcher2 consists of the values 4, 6, 8, and 10. The hash table stores the intersection of the values 4 and 8. The result of this step [17] is a collection of intersections or subspaces. The crossings or subspaces in S are organized in a decreasing order according to magnitude. For each row in dataset S, assess each subspace in respect to all of its succeeding subspaces. If a subspace is a subset or equal to subspace si, then remove it from the list of subspaces. Compute the extent of similarity or intersection between the given subspace and the other subspaces in the list of subspaces S for the purpose of forming clusters. The threshold parameter is employed to control the degree of overlap, representing the proportion of dimensions/elements that concur. If the similarity between the two subspaces surpasses the designated threshold, then the subspace is selected as a constituent of the cluster. This method is repeated for each member in the list of subspaces, resulting in a significant number of clusters of subspaces. After the researcher logs in, the system gets the papers that have been given a rating. The subspaces containing the papers of the current researcher are collected, and the matching subspaces, denoted as [18], are retrieved from S. The acquired articles or subspaces are ranked based on their resemblance to the researcher's current selection. The researcher suggests that the active user explore the articles in the top-ranked subspaces that have not yet been rated or read. The other related references are [19-22].

5.1Research Paper Recommendation Algorithm

Researcher-paper matrix accompanied with a roster of study domains. The result is to compile an inventory of papers and select the field of research. Aggregate the catalogue of academic publications in the field of research. The following steps have been performed:

Step 1: Prepare a researcher-paper matrix as per the given subject. Perform the specified action for each row, labelled as r1, in the matrix. Transform r1 into a string including identified numbers of articles that have been awarded a high grade (4 or above) by the researcher;

Step 2: Set the hash-table h to a null value. Iterate through each row, represented as r1, in the altered matrix. Next, do a comparison between the first row, referred to as r1, and the following row, labelled as r2. If there is a common part between the sets r1 and r2, then add this common part to the set h, along with the researcher's identity, and increase the count. Establish a starting value for the threshold. Include intersection i in subspace list s if its size is greater than or equal to the threshold, for every intersection i in h.

Step 3: Organize the components in s in a descending order according to their magnitude. Compare subspace s1 with following subspace s2 within the set s. If s2 is a subset or equal to s1, then delete s2 from s.

Step 4: Compile the articles which have high ratings by the active user and label this group as "q". If there exists a subspace s1 that is a subset of s, and the intersection between s1 and q is equal to or larger than the threshold, then s1 is selected as a member of cluster c.

Step 5: Organize the members of c based on their extent of addressing the query. The subspaces with higher rank are chosen.

6. Results and Discussion

For computation of the results, simulated dataset is created that consists of 30 unique areas of study. Each area includes more than 300 research papers written by over 200 scholars. The system is tested by generating and inserting new research articles. The technology demonstrated robust performance since it is independent of the number of dimensions. Clusters were formed completely in all cases, without any further clusters being documented. If the new user just reads and evaluates one paper, the quality of recommendations will be reduced. Nevertheless, it is rare to come across such a circumstance, as most researchers have a comprehensive grasp of the subject and are aware of which papers to refer to. As a result, the overall quality of the system will not be impacted. Precision and recall are often used measures for evaluating the quality. These two factors are observed through following formulae

$$\text{Precision} = NP/NR \quad (1)$$

where, NP=Number of relevant paper and NR=Total number of retrieved papers.

$$\text{Recall} = NP/NCR \quad (2)$$

Where NCR = Number of relevant papers of the database.

Table 1. Computation of Precision and Recall Factors

Sno	NR	NP	NCR	precision	recall
1	40	37	38	0.925	0.973
2	40	38	39	0.926	0.974
3	150	133	147	0.886	0.904
4	350	306	345	0.874	0.886
5	600	506	621	0.843	0.814
6	1050	860	1052	0.819	0.817

The values of precision are calculated as
(1) precision=37/40=0.925
(2) precision=38/40=0.926
(3) precision=133/150=0.886
(4) precision=306/350=0.874
(5) precision=506/600=0.843
(6)precision=860/1050=0.819
The values of recall are calculated as
(1) Recall=37/38=0.973
(2) Recall=38/39=0.974
(3) Recall=133/147=0.904
(4) Recall=306/345=0.886
(5) Recall=506/621=0.814
(6) Recall=860/1052=0.817

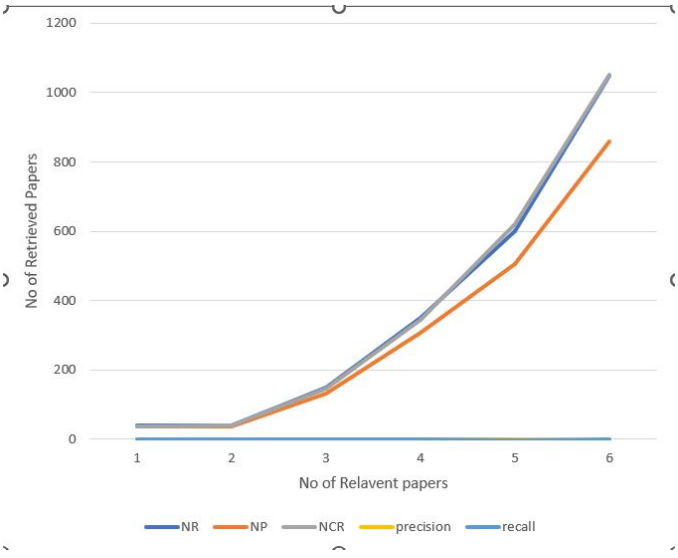

Figure 3. Precision Versus Recall Factors

The number of articles examined by the active researcher has an impact on the accuracy and recall measures. Enhanced precision may be attained by the researcher while reviewing a greater quantity of publications, rather than assessing only a sample of articles. Let NR be the overall number of researchers, NP denotes the entire number of papers, and NCR denotes the number of articles evaluated by the current researcher. Experimental analysis is conducted to assess the efficacy of suggestions by modifying the variables NR, NP, and NCR. When researchers are increasing, then the quality of ideas controlled by NCR is also increasing as presented in table 1 and figure 3.

8. Conclusions

The result of this investigation suggests for employing subspace clustering as an effective method to market research articles to researchers. Subspace clustering specifically targets significant dimensions, hence improving the efficiency of the recommendation process. Research articles are recommended based on the assessments supplied by specialists in the corresponding field. Consequently, this provides excellent suggestions and functions with exceptional efficiency. A thorough elucidation of the creation and application of subspaces are given to improve the excellence of present system by suggesting scholarly papers.

Disclaimer (Artificial intelligence)

Author(s) hereby declares that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during writing or editing of manuscripts.

References

[1] Akhtar, N. and Agarwal,Devendera. “A Literature Review of Empirical Studies of Recommendation Systems”, International Journal of Applied Information Systems,USA , Volume 10, No. 2, Pages 6 – 14, December 2015.
doi:10.5120/ijais2015451467

[2] Dong, Li-yan, Wang, Yu, Ren, Yi and Yong-li Li, "Collaborative Filter Algorithm Based on Matrix Decomposition and Clustering", Journal of Jilin University (Science Edition), vol. 57, no. 01, pp. 105-110, 2019.

[3] Zhu, P., Zhu, We and Hu, Q., Zhang, C. and Zuo, W., "Subspace Clustering Guided Unsupervised Feature Selection", Pattern Recognition, vol. 66, no. C, pp. 364-374, 2017.
https://doi.org/10.1016/j.patcog.2017.01.016

[4] Calma, A. and Davies, M., “Studies in Higher Education 1976–2013 : A Retrospective using Citation Network analysis,” Studies in Higher Education, vol. 40, no. 1, pp. 4–21, 2013.
https://doi.org/10.1080/03075079.2014.977858

[5] Garfield, E., “Citation Analysis as a Tool in Journal Evaluation.”, Science, vol. 178, no. 60, pp. 471–479, 1972.
doi: 10.1126/science.178.4060.47

[6] MacRoberts, M.H. and MacRoberts, B.R., “Problems of Citation Analysis: A Critical Review,” Journal for American Society of Information Science, vol. 40, no. 5, pp. 342–349, Sep. 1989.
https://doi.org/10.1002/(SICI)1097-4571(198909)40:5<342::AID-ASI7>3.0.CO;2-U

[7] MacRoberts, M.H. and MacRoberts, B.R. “Problems of Citation Analysis,” Scientometrics, vol. 36, no. 3, pp. 435–444, 1996.
https://doi.org/10.1007/BF02129604

[8] Beel, J., Langer, S., Genzmehr,M.,Gipp, B., Breitinger, C. and Nürnberger, A., “Research Paper Recommender System Evaluation: A Quantitative Literature Survey,” RepSys, vol. 20, no. April, pp. 1–35, 2013.
doi:10.1007/s00799-015-0156-0

[9] Beel, J.,Gipp, B., Langer, S. and Breitinger, C., “Research-Paper Recommender Systems: A Literature Survey,” International Journal of Digital Libraries, vol. 17, pp. 305-338,June, 2015.
https://doi.org/10.1007/s00799-015-0156-0

[10] Beel, J., “Towards Effective Research-Paper Recommender Systems and User Modeling based on Mind Maps,”, A Dissertation, 2015.
Towards Effective Research-Paper Recommender Systems and User Modeling based on Mind Maps (beel.org), Last Visited 31.05.2024

[11] Akhtar, N. and Agarwal,Devendera “An Efficient Mining for Recommendation System for Academics”, International Journal of Recent Technology and Engineering, Volume-8, Issue-5, Pages 1619-1626, 2020.
doi: 10.35940/ijrte.E5924.018520

[12] Jain, A. and Vishwakarma, S.K., "Collaborative Filtering for Movie Recommendation using RapidMiner", International Journal of Computer Applications, vol. 169, no. 6, pp. 0975-8887, July 2017
doi:10.5120/IJCA2017914771

[13] Xue, G., Zhang, H., Bian, J. et al., "Learning Image and User Features for Recommendation in Social Networks[C]", IEEE International Conference on Computer Vision IEEE, pp. 23-36, 2015.
Learning Image and User Features for Recommendation in Social Networks (cv-foundation.org) Last Visited 31.05.2024

[14] Tierney, S., Gao, J.B. and Guo,Y.. "Subspace Clustering for Sequential Data", Proc. Of IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1013-1020, 2014.
doi:10.1109/CVPR.2014.134

[15] Peng, C., Kang, Z. andCheng, Q., "Subspace Clustering via Variance Regularized Ridge Regression", Proc. IEEE Conference Computer Vis. Pattern Recognition, pp. 2931-2940, 2017.
doi**:** 10.1109/CVPR.2017.80

[16] Tsai, C.F. and Hung, C. "Cluster ensembles in collaborative filtering recommendation", Applied Soft Computing, vol. 12, no. 4, pp. 1417-1425, Apr. 2012. https://doi.org/10.1016/j.asoc.2011.11.016

[17] Perwej, A., Perwej, Y., Akhtar, N. and Parwej, F "A FLANN and RBF with PSO Viewpoint to Identify a Model for Competent Forecasting Bombay Stock Exchange", COMPUSOFT, SCOPUS, International Journal of Advanced Computer Technology, 4 (1), Volume-IV, Issue-I, Pages 1454-1461, 2015. doi: 10.6084/ijact.v4i1.60

[18] Maazouzi, F., Zarzour, H. and Jararweh, Y., "An Effective Recommender System Based on Clustering Technique for TED Talks", International Journal of Information Technology and Web Engineering, vol. 15, no. 1, pp. 35-51, Jan 2020. doi: 10.4018/IJITWE.2020010103

[19] Roy,D. and Dutta, M., "A Systematic Review and Research Perspective on Recommender Systems", Journal of Big Data, **9**, 59. 2022. https://doi.org/10.1186/s40537-022-00592-5

[20] Singh, R., Gaonkar, G., Bandre, V., Sarang, N. and Deshpande, S., "Scientific Paper Recommendation System," 2023 *IEEE 8th International Conference for Convergence in Technology (I2CT)*, Lonavla, India, pp. 1-4, 2023 doi: 10.1109/I2CT57861.2023.10126196.

[21] Saxena, V., Verma, V., Verma, V., Singh, K.V. (2024). Data Cube Technology for Accessing of Large Database. In: Devi, B.R., Kumar, K., Raju, M., Raju, K.S., Sellathurai, M. (eds) Proceedings of Fifth International Conference on Computer and Communication Technologies. IC3T 2023. Lecture Notes in Networks and Systems, vol 897. Springer, Singapore. https://doi.org/10.1007/978-981-99-9704-6_4

[22] Choudhary, B. and Saxena, V., "Faster Access of Credit Cards through Data Cube Technology", 2021, Webology, Vol. 18(6), 4920-4933. https://www.webology.org/