Literature Recommendation System to recommend the best true research papers to the user

Akshata Saptasagar
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
akshata.saptasagar@gmail.com

Sandhya Waghere
Faculty of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
sandhya.shindhe@pccoepune.org

Rahul Badgujar
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
rahulbadgujar1508@gmail.com

Atharva Misal
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
asmisal26@gmail.com

Omkar Raskar
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
omkarmraskar@gmail.com

***Abstract*— There are various domains where recommendation systems are modelled and one of them is the Literature-based recommendation system approach. This paper uses a hybrid approach of recommendation systems to recommend articles. This paper proposes a model that compares time-complexities of various techniques to find the most suitable compatible algorithm which is compatible with the user's system and recommends the article using the algorithm that has the least time complexity. This system also considers the truth value of the research papers. The proposed model provides the top-recommended article along with a reference list of citations and organizational ratings.**

***Keywords*– Recommendations System, Literature, hybrid approach, time-complexity, citations.**

I. INTRODUCTION

Research papers are types of academic literature survey papers that provide independent interpretive in-depth analysis on a user-specific topic. Over the years, millions of research papers are published in various journals and are made available to the world via the world wide web. These research papers are recommended to the user via a Literature-based recommendation system. Recommendation systems are the models which aim to suggest relevant items to the user based on their most recent searches on the internet. There are 7 approaches for recommendation system development which include: Content-based recommender system (CBRS), Collaborative filtering recommender system (CFRS), Demographic recommendation system (DRS), Hybrid recommender system (HRS), Knowledge-based recommender system (KBRS) and Context-aware recommender system (CARS).

A. Principle of Recommendation System

Content-based Recommendation Systems are the keywords of the system that are allocated to specific contents. While the user is surfing on the web, the user's rated as well as surfed content is taken into consideration and the system recommends similar items to the user. Collaborative Filtering recommendation systems work on assumption principles where commonalities are taken into consideration and user-specific recommendations are provided. This is one of the most efficient and used algorithms in Recommendation systems In Demographic Recommendation systems data is collected and categorised into specific classes. Using person-to-person correlations, items are recommended to the user. In this system, user rating history is not considered. A hybrid recommendation system includes combining two recommendation systems to make an industry-specific model. Because of combinations, limitations of the systems are overcome and strengths are mixed, giving rise to a powerful yet industrial efficient model. Knowledge-based recommendation systems work on market-basket analysis methods. This system interprets the user's choices and based on these, specific items are recommended. In utility-based recommendation systems, various industries use various object considerations to meet user-specific utility requirements. This system is used to build E-commerce recommendation systems where people can check object availability, real-time inventories and many more.

II. RESEARCH AREA

In the present paper, we are going to propose a model that is based on a hybrid recommendation approach. This recommendation system will recommend research articles to the client based on the client's interest, rating, history and system capabilities. The paper focuses on the systems as well as users, where systems will be able to use more than one recommendation approach and time complexity will be taken into consideration and the algorithm which will be using the The least time to execute will be used to implement the necessary recommendation actions. While searching these papers there are possibilities that a person might encounter a situation where he/she has been provided with several research papers based on his/her choices. Also, there are possibilities that the provided paper might be falsely published and made available on the internet. so, our paper will not only focus on time complexity but will focus on providing research papers that are truly published by considering genuine publication sites, paper citations and organizational ratings.

III. RELATED WORK

A. User Interest and Citations Based Paper Recommendation System

In this paper, Betül Bulut et al. had proposed a recommendation system approach that is specific to the user profile. Their proposed method tries to determine the researcher's level of interest-based on the publications he or she has produced. Their approach concentrates on the researcher's paper bibliographies and citations. This study is a bibliography-based approach that focuses on the user profile. [1]. The author put forth a content-based recommendation system for the purpose of filtering articles. But solely content-based recommendation does not give a satisfactory result. Research done on the hybrid approach [2] shows that the accuracy of recommendations is increased and the system achieves better performance. To make research' work simpler by shortening the time they waste looking for articles is the primary objective of this paper.

In this model, before applying any algorithm, obtaining datasets, preprocessing on that and putting together a user profile is done. The proposed recommendation model uses two algorithms for gathering data and recommending it to the user. The clustering of the bibliography is finished using the K-Means (K) must be created within the process. To figure out which cluster is the most similar to the user's profile K-NN Algorithm is employed. The K-nearest neighbours (KNN) algorithm predicts the values of recent data points based on 'feature similarity'.

The proposed model uses the K-means Algorithm for clustering of article bibliographies and the K-NN algorithm to find the recommended result. It was compared to the F1 metric best N recommendations precision, recall, and. To integrate these two measurements with equal weights, the F1 metric was utilised as a harmonic mean concerning recall as well as precision.

According to the experimental observations, the suggested model's accuracy falls as the figure of Top-N predictions along with recall and F1 score increases. The accuracy was higher when selecting the Top 10 items when the F1 measure was higher, according to the findings of this study.

The model recommends personalized articles based on user profiles and bibliography. Based on the experimental test conducted in the paper, it is concluded that the proposed model gives better accuracy to recommend the top 10 articles with high F1 scores. Filtering candidate articles by a year helped in increasing the accuracy of recommendations.

B. Recommendation of Academic Papers using Heterogeneous Information Networks

Progress in the fields of science and technology are made at much greater speed but to keep a record of this progress, research papers are published under various organizations. Over the decades, millions and millions of research papers have been published and made available to the world. To search these papers and to acquire appropriate information, most people across the globe make use of Google or Google Scholar. The existing methodologies for research paper recommendations are based on the keywords specified in the research paper, or authors or the year of the publication. But using this method, most people don't find appropriate literature papers. The recommendation system consists of mainly three recommending algorithms. They are Collaborative Filtering, Content-based learning and Graph-based Filtering. A method where researchers who possess an interest in similar topics would be able to publish as well as search for the papers is called Collaborative Filtering. A particular Network (HIN) is another name for the Graph-based Filtering technique. In this technique, citation relationships, collaborating and contributing relation- ships and research regions are presented in graphical format. Many times synonyms of most of the words are used to avoid plagiarism.

In this paper, the authors have proposed to construct a HIN where the recommendation model will be able to detect similar worlds and recommend papers that consist of highly accurate information. To check for applicability, experiments were conducted on large public digital libraries. This experiment overcomes the limitations of one-sidedness in collaborative filtering. Also, paper quality is taken into consideration which was one of the limitations of the Content-based filtration technique. Various pieces of information like the collaboration of authors, citations included research papers and research regions are included to build two categories of HIN. A strategy naming random walk is used to conduct traversals across edges and models naming natural language are employed to match sequences of words.

- HNPR Algorithms: Two heterogeneous information networks were constructed, namely: Network of Paper along with Author and Network of Paper along with area.
  - Network of Paper along with author: Three edges are named: the first one is the undirected edge which denotes the authors of literature work. Secondly, there are a set of edges that are undirected and denote that the authors or co-authors have written one or more papers. The third one is the

set of directed edges where citation reference is denoted.

- Network of Paper along with area: Authors need to manually select the research regions provided by the system. By inserting proper terminologies, the system will be able to provide a list of regions where the word is located.

- Random Walk: Natural Language models help in determining two similar words without using any representation techniques. By using front and back Random walk sequences, papers are compared and if similarities arrive, then the two papers are concluded to have similar contents.[4]
- Natural Language Model: The citation relationships among different papers and the contributions made by the authors are measured in the network. named paper-author. In the network named paper-area, information regarding research regions is extracted.

5000 articles were considered and papers with alike references regarding particular patterns were sorted. Using Performance Metrics, the system was able to sort genuine papers and also recommend new papers which consisted of relevant information[3]. In this paper, a method was proposed naming HNPR, by considering contextual and structural information presented in papers. The system recommends more than 30-percent of accurate papers compared to the HINE which is a classical graph-based algorithm.

C. A content-based recommendation perspective constructed on LDA modelling for article recommendation

Recently, a lot of research is being conducted to create a recommender system for recommending literature. Existing approaches are used in the models proposed for Recommendation based on Content. Deep Learning and Topic Modelling are some novel approaches also being used.. One of the major give accurate recommendations over the diverse compilations. Few failed to capture the complexities around current relationships which exist in Corpus. Mr Dhiraj Vaibhav Bagul and Dr Sunita Barve proposed a content-based model of literature recommenda- tion using LDA (Latent Dirichlet Allocation) for literature recommendation purpose. The suggested model also uses Jensen-Shannon distance method to calculated the relevance score among documents. The model is targeted to resolve the issue of Scalability and to perform well in recommending literatures over a broad level of domains.

The proposed model in this paper has three varied stages of algorithms. The initial step is document cleaning, document preparation and vector representation. After that, a topic model is created which is based on LDA, which works undergoing a process where the words appearing in the document provide essential data about the semantics of the document, and expects that the document relevant to it also contains the set of words. Finally, a recommendation module based on Jensen-Shannon distance [5] will produce recommendations, based on the distance score. The lower the score, the higher the similarity between the two documents.

Different Natural Language Processing techniques are employed in the recommender model proposed in this paper to perform preliminary processing of the text and strategies for feature weighing in order to produce a model which is vectored for all the documents. When all the ceasing words from the document are removed, the algorithm for stemming given by Martin F. Porter is incorporated to trunk the words. To create a vector depiction for the documents, frequency-based embedding techniques are applied. In the evaluation, the counter-vector approach outperforms TF-IDF Vectorization, hence it is used in the proposed model. In order to classify the literature documents on proportional topics, the LDA topic modelling technique [6] is used. Previously published research papers metadata is used to build the Topic Model based on LDA. To carry out this genism, a python library is used. Similarity between the documents is identified using the similarity score computed using The Jensen- Shannon distance [5] to recommend top-10 documents relevant to a given manuscript in the ascending order of relevance. While evaluating the model, a python library arcas is used to collect metadata of 500 papers for each of the publications.

The recommender model proposed in this paper can recommend top-10 publications based on the content similarity with the manuscript abstract given by the user as input. The model was evaluated using the compilation of 100 publications belonging to 10 different domains. Abstract and suplementary metadata of these research papers which were published previously reflect the contents of these publications, hence they were used. The suggested LDA Model based on count-vector illustration outperformed the LDA Model based on TF-IDF vector based model and also the cosine similarity models. The model has achieved the precision score of 62.58-percent and 68-percent for unseen and seen data respectively in recommending top-10 relevant suggestions. The model outperforms the cosine-similarity-based ap- proach in capturing subtle topical relationships across a large corpus. As a result, we can consider applying this model for literature recommendation tasks, where the system generalise with the time for the documents distributed amid a broad along with varied collections to give relevant recommenda- tions for seen as well as unseen documents.

D. Graph Structure and user interest based Paper Recommendation System

The main purpose of recommendation systems is to provide suggestions to the users based on previous data. Many research papers are published till today, and so it becomes difficult to search for them. This problem is known as information overload [7]. The users have to invest a lot of time searching for the right research paper. Hence it has become essential to develop a recommendation system that will recommend the correct research paper that the user wishes to seek. In existence, few recommendation systems help with the purpose but their scope is limited. Although the Collaborative Filtering method was used, its performance was not satisfactory. Hence the use of the auxiliary method was adopted to increase the performance of the recommendation system.

The proposed model includes citation structure being the first recommendation system type which then searches for similar papers and displays the result. Since all the required work cannot be found in one paper, therefore a reference list with top-N relevant papers is involved. Further, a paper graph is constructed with the help of reference citations. In some cases previous papers cannot refer to recently published papers, some information can be missing or a few irrelevant papers can be adopted. Hence, there's a network called Memory Network (MN) that helps inefficient handling of the sequential data and memorizing long-term dependencies. This has been added to the recommendation systems.

As the author is more interested in recently searched papers hence he/she is more precise to interact regarding those papers. So it is of utmost importance to recommend the papers sequentially following the author's interest. Further, there's an extraction of —L— in successive papers for every author u. Then a graph to seize the relation between the papers is set up as they aren't inherently graphs. This is done for all the authors and counts how many edges of the paper which was and nearer to each other in order. For the latter part, a GNN is used that aggregates the closer papers in Lu, Ll to know about the interest of user in paper representation. Then for every graph, its content similarity will be calculated by each node.

This method is more preferable to the traditional methods. Methods like LDA-TM [7] do not consider the scenario about the paper citation or supplementary information sources. Other methods like the CDL and ConvMF do not consider the user's interests and cannot provide a useful recommendation. The MEIRec focus it's large part on structural information whereas TAAS emphasizes textual information. Hence this method is more suitable.

This paper showed a way that referred to the interest of the user periodically and network of the graph structure and has obtained great information to recommend papers. Here the user's interest and calculated structural attention has created a graph structure where the user's interest was taken into account.

IV. PROPOSED WORK

Literature-based Recommendation Systems mostly include algorithms that rarely tackle the problems regarding time complexities. The algorithms used to implement the system mostly consider citations and authors interest irrespective of system capabilities and time complexities. The proposed model is a Hybrid Recommendation System that focuses on recommending true papers to the user by keeping into account the matter of time complexities. The user's interest, system capabilities, paper citations, genuine publications and algorithms will be taken into consideration to build a whole new literature recommendation system.

V. CONCLUSIONS AND FUTURE SCOPE

As a hybrid recommendation model is a combination of two or more recommendation systems, so one such algorithm will be developed based on this hybrid RS, where two or more algorithms will be compared and evaluation will be done based on the author's interest, citation, system capability and time complexity.

This paper involves a model of a recommendation system where a Literature RS is developed using a hybrid algorithmic approach. We developed a model which considers time complexity and recommends genuine papers to the user.

REFERENCES

[1] B. Bulut, B. Kaya and M. Kaya, ”A Paper Recommendation System Based on User Interest and Citations,” 2019 1st International Informatics and Software Engineering Conference (UBMYK), 2019, pp. 1-5, doi: 10.1109/UBMYK48245.2019.8965533.

[2] A. Tsolakidis, E. Triperina, C. Sgouropoulou, and N. Christidis, “Research Publication Recommendation System based on a Hybrid Approach,” Proc. of the 20th Pan-Hellenic Conference on Informatics, 2017.

[3] X. Cai, J. Han, S. Pan, and L. Yang. Heterogeneous information network embedding based personalized query focused astronomy reference paper recommendation. Int. J. of Computational Intelligence Systems, 52(1):591–599, Mar. 2018.

[4] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In Proc. of the 20th ACM Int. Conf. on Knowledge Discovery and Data Mining, pages 701–710, Mar. 2014.

[5] CD Manning, H. Schu¨tze, “Foundations of statistical natural language processing,” MIT press; May 1999.

[6] DM Blei, AY Ng, MI. Jordan, “Latent dirichlet allocation,” Journal of machine learning research, Jan 2003, 993-1022.

[7] Sugiyama, K., and Kan, M.-Y. 2010. Scholarly paper recommendation via user's recent research interests. In Proceedings of the 10th Annual Joint Conference on Digital Libraries(JCDL), 29–38.

[8] D. V. Bagul and S. Barve, ”A novel content-based recommendation approach based on LDA topic modeling for literature recommendation,” 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 954-961, doi: 10.1109/ICICT50816.2021.9358561.

[9] N. Du, J. Guo, C. Q. Wu, A. Hou, Z. Zhao and D. Gan, ”Recommendation of Academic Papers based on Heterogeneous Information Networks,” 2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA), 2020, pp. 1-6, doi: 10.1109/AICCSA50499.2020.9316516.

[10] H. L, S. Liu and L. Pan, ”Paper Recommendation Based on Author-paper Interest and Graph Structure,” 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2021, pp. 256-261, doi: 10.1109/CSCWD49262.2021.9437743.