International Journal of Scientific Research & Engineering Trends
Volume 5, Issue4, July-Aug-2019, ISSN (Online): 2395-566X

Tag Based Recommender System for Scholarly Data

Mohit Kumar Prof. Shilpa Verma
Department of Computer Science and Engineering
Punjab Engineering College
Chandigarh, PB, India

***Abstract*- Recommender Systems are being used nowadays in every field online and they help to play an important role in user/customer satisfaction and attraction. The motivation to write this article comes from the struggles that a research scholar goes through while finding the perfect paper to read. In this article, the aim is to generate recommendations to the users by eliminating the cold start problem. This paper works on finding a suitable article for the user by keyword matching from user profile and existing data set. Various different bibliometrics have been used in this article to apply the underlying method for generating a similarity index for the documents. The method of tf-idf score calculation has been used to find the similarity between the keywords the user is interested in and the keywords that occur in the content of a certain article. The scores were calculated for each document by finding the relevance between the user keywords and paper keywords to find level of relevance of the documents and hence generate the recommendations accordingly.**

***Keywords*- Tag Based recommender system, Content based, Collaborative filtering, Hybrid filtering, Cold Start Problem.**

I. INTRODUCTION

Recommender systems are slowly becoming part and parcel of the online world day by day. We can see the use of recommender systems almost everywhere from online shopping sites to the movies and web series recommendations. As we are familiar with the rapid growth that has been happening in today's world , also the technologies and techniques are growing rapidly. As a result of this, there has been an increase in people opting to do a research over the new and old techniques and make a contribution to their respective fields.

Now to actually be able to do a research, a user will need to do a search first and they will need to have something related to their field. How to get that? The simple answer to that is to study research papers, journals, conference papers etc. Now as there is a growth of researchers and there work, so it also contributes to the growth of number of conference papers in every field. So, to help the researchers find a suitable conference paper, we study the recommender systems and try to develop one using the existing filtering techniques.

Here, recommender systems are used to generate the recommendations of papers to the research scholars. A Recommender System is similar to a shopkeeper who will show you items that you have the highest chances to like and purchase based upon various different factors such as your past purchase history, your attraction towards a particular type of product, or based on the most liked items by majority of customers. These ideas or approaches are categorized into content based and collaborative filtering. In this paper, a tag based recommender system has been developed which will extract the author paper keywords and match them with tags from user profile to generate a score. Higher score means better suitable article and hence more chances of paper being recommended. This whole process is basically carried out in different iterations where user tags are matched with keywords from paper title, abstract and author paper keywords. To provide the diversity in recommendations, a second algorithm will be used which will take into consideration the bibliometric measures such as number of citations and references in the paper to generate similar recommendations.

II. LITERATURE SURVEY

1. Tag Based Recommender Systems

By tagging, we understand that some particular keywords are being assigned to some content. Tag based recommender systems the ones in which the digital content is tagged with relevant keywords, hence dividing the data into different categories. This helps in recommendation to the users as well searching of data by users merely based on specific keywords of interest. There are several different categories into which tag based recommender systems have been classified based on the type of content been tagged.

2. Application Oriented Recommender Systems

The following table shows the use of recommender systems in various different fields of application. These all are related to the research articles but the types of recommendations generated are of different fields e.g medical articles, citing systems, classical research papers etc. These all have various types of techniques of recommender systems used in them such as content based, collaborative filtering etc.

1. Folksonomy - This includes the tags from an information service.
2. Personomy- This includes the tags given by a person
3. Docsonomy- This includes the tags of a document that is concrete.
4. Joursonomy- This includes the tags from a concrete journal in some Social bookmarking service.
5. Tweetonomy- This includes the hash tags used on twitter.

The survey conducted on the tag based recommender systems included various papers from different conferences, journals. These papers belong to a wide variety of fields of application and involve use of different approaches. We have majorly focused upon the folksonomy and personomy types of tagging systems. Following section includes a brief discussion about all the papers studied.

3. Folksonomy Based Recommender Systems

Content based recommender systems are generally based on analyzing the content and generating the recommendation results. The authors in [ollagnier et al., 2018] have focused on analysis of the content and coined a main term called 'Centrality Indicator' which is used to evaluate the importance of bibliographic references of a paper that the user is reading. The main drawback of this approach was the high complexity of topics and hence there was a dif-ficulty in evaluation of the papers.

A large amount of user activity can be seen on social networks these days. One such use of this type of data from social networks has been used in [Bahulikar, 2017]. The main approach used by the author is content based filtering and this paper has used the social networking based tags used by a user and also the location based services of the user in order to generate the appropriate recommendations for the users. The frequency of accessing the social media account by a user was collected and weights were calculated accordingly.

According to A.Wijonarko et al. [Wijonarko et al., 2017], three different methods i.e Random Walk with Restart, content based, collaborative filtering were studied and each was applied on the Social Tagging System. Then a new hybrid method was generated by combining all these three techniques and the results of hybrid method were compared with the three individual methods. Hybrid method produced the best results among these four methods. The use of Natural Language Processing in the hybrid method to cluster the tags having same meaning could prove to improve the accuracy of the system.Recommender systems have a wide variety of applications online. From movies to shopping, there is another field of application known as arts commission where artwork of various artists can be recommended to put up at display.

In image recommender for art commissioning [Kosala et al., 2017], the recommender system is made to generate recommendations of art- work images for an arts commission. The system created by author involves the use of content based approach and makes the use of features such as histograms, contents, category, descriptive tags, artwork image, user-generated tag contents, art style category etc. The accuracy of the system was tested not by the algorithm but by directly taking the reviews of the buyers about the usefulness of recommendations made to them. Reference management tools widely used for creating and managing the citations in a research paper. These tools are a boon for the research scholars.

In paper [Kaur and Dhindsa, 2016], the reference management software are being reviewed. Reference management software basically helps the authors to generate bibliographic references to the papers being cited in their own research article. In this paper, the author has compared three different reference management software and carried out various differences among them. All the three have some advantages and disadvantages when compared to each other. In the end, the author has conveyed that it all comes to a user preference level.High-Utility Item set Mining (HUIM) is an important data mining tool in transactional databases Efficient Incremental High-utility Item set tries to improve the efficiency of HUIM.

In paper [Dhanda and Verma, 2016], a previously proposed existing technique known as Efficient Incremental High-Utility Item set Mining algorithm (EIHI) has been used as a base and a few modifications have been applied to this method. The main reason for using this method as base is because it can work on dynamic data sets. Two main modifications are added over the existing method. first one includes in finding the papers based on its content that is appropriate for the user. Second one includes to find the reference sets having high utility for a user based upon there personalized requirements.

According to work done by A. S. Tewari et al.[Tewari et al., 2016], the use of collaborative filtering has been deployed onto the users as well as items and then their respective matrices are generated. After that the similarity index is found for user-user and item-item types which provides an average value that can be later referred to generate recommendations. It may suffer from cold start problem if tags given by the user are very few or non or say there is a new item for which similarity index is yet to be found.Content based and collaborative filtering techniques of recommender systems have both advantages' and disadvantages over each other. A hybrid system can be used to combine the merits of both the systems and can be helpful to overcome the drawbacks of

the two systems. In one such research article [Kumar, 2015], the authors have used combination of content filtering and collaborative tagging and focused on removing the cold start problem. The system allowed the users to upload their own articles and also give them tags manually. This freedom of assigning tags manually can lead to reduced efficiency of system if a user provides wrong tag for an article.There is always a confusion because of which people often tend to treat keywords and tags the same thing. In reality, they both are different entities. Keywords are a part of the content and used to identify what is there in content. Whereas tags are placed by the content creator.

In the work done by D.Anand [Anand, 2014], the efficiency of keyword based system is com- pared with the tag based system and merits and demerits of both have been displayed. Overall, keyword based being the better one structurally and expert monitored still fails to match the accuracy of tag based system. A total of three factors i.e precision, recall, F1-measure were taken into consideration to compare the accuracy of both systems. One common practice in collaborative systems is to generate the trends from the top most rated items or say topics.

According to this assumption, a list of most popular keywords has been created in [Zhou, 2012] from year 2007-20011 and used to generate the hot topics in the given time period. The use of E-utilities has been made in order to collect the number of publications in PubMed and hence extract the keywords from these publications. The main drawback in this paper is that the keywords can also be assigned by the users themselves. A user can assign wrong tags or keywords to some topics and hence it can affect the efficiency of the system.

In paper [Sen and Riedl, 2011], an attempt has been made on formation of folksonomies by generating various types of tags. Folksonomy is nothing but categorization of digital content based on different tags. Major types of user tags discussed are shared, unshared, shared-pop, shared-rec. There is a major challenge in this paper which is how to decide what makes one tag more useful than others. Traditional collaborative systems may be slow to produce relevant results for a new user and may also suffer from the cold start problem.In the works of C.

Hang et al. [Hang and Meifang, 2011] , a tag-semantic similarity method has been used in order to improve the result that are produced by the traditional collaborative filtering system. The main drawback of using traditional collaborative filtering is that it does not take into account the user preferences and hence this results in decreased efficiency. To weight the quality of item tags TF-IDF method has been used in this paper. Although, the accuracy achieved in this system is still very less and the relevance of tags still needs to be improved. The use of content based and collaborative filtering techniques alone cannot be sufficient enough to produce efficient results.

According to the author, in [Alepidou et al., 2011] , the hy- brid techniques of recommender system should be preferred to using only a single one. Training the algorithms based on semantic matching criteria may result in more relevant recommendations. The main limitation of this method is that the determination of best overall evaluation criterion is not achieved and hence it can be worked upon in future.

Similarly in the work done by A Kohi et al. [Kohi et al., 2011] , the use of hybrid methods is stated to be more accurate and relevant as compared to individual methods. Also the method used here can be used for large real world problems due to its scalability. It has a simple and fast implementation. The major drawback of this paper is that semantic relations among users should be utilized and social networks among users need to be discovered in order to create a better output. A major issue of the tag based systems is that there is no hierarchy representing relationships among tags and hence it is very difficult to build the structure for tags.

In paper [Yoo and Suh, 2010] , this issue is solved using a new User-Categorization of tags. In this method, the tags are categorized i.e UC Tag consists of tags and their tag categories. The main limitation is that the system has only one type of relationship i.e. has relationship. There is a need to add other relationships such as: part-of, same-as, and inverse-of etc.

III. PROBLEM STATEMENT

Cold start is a big problem in the field of recommender systems. It means that it is difficult for a recommender system to generate relevant recommendations for a new user as there is not much information available about the user. Here we aim to design and develop a tag based recommender system using content based filtering to analyze academic social networks using scholarly data and overcome the cold start problem.

IV. WORK DONE

The main problem identified in many recommender systems has been the cold start problem. We aim to solve this problem by using the keywords provided by the user as it will create an idea of the choice of user's likings and hence help us in generating the relative results for a new user. The main criteria used for implementation of this whole technique are the classic tf-idf approach. Here tf stands for term frequency and idf stands for inverse document frequency. The main idea of the tf-idf approach is to find out the importance of a word inside a document based upon the number of times a word appears inside the document with respect to the number of times that word

appears inside the complete data set. An algorithm has been designed for the recommendation of the papers from the data set. This algorithm tries to remove the cold start problem as much as possible. The following is a brief discussion of the algorithm followed by the pseudo code.

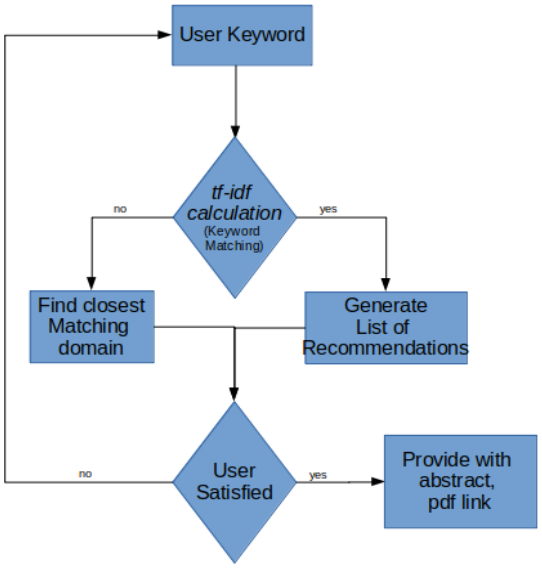

Fig.1 Methodology flow control.

1. Dataset

We have focused on getting the data from a single source for starting this work and our main target has been iee explore.org. All the documents that were used as part of dataset were collected from ieeexplore.org. Details like document title, publication title, year, citation count, reference count, pdf link, author keywords, ieee keywords etc are extracted and used for the keyword matching purpose. Pre-processing techniques (tokenization, stop word removal etc.) were applied on the data to extract the required keywords.

2. Data Structures used

Python lists have been used to store the set of documents

- Python dictionaries have been used to store the keywords and their frequency counts.
- D s is a list in which each element is the document weight and is initialized to 0.
- D t is a list of lists. Each list element contains the list having title keywords.
- D u is a list of tags extracted from the user profile.
- D a is a list of lists. Each list element contains a list having the keywords extracted from abstract of a document.
- D p is a list that contains the author paper keywords. Each element in this list is another list which has the keywords that belong to a certain document.
- The functions update t , update a , update p are used to add a predefined score to the old weight of a document.
- These scores are 0.5*(frequency of keyword), 0.2*(frequency of keyword) and 0.3*(frequency of keyword) respectively. This denotes that a keyword appearing in the title has more relevance than a keyword appearing in the abstract and author paper keyword lists.
- The function append is used to to insert the element into a list.
- i, j, k all are loop variables.
- C is a list containing the number of citations of each document.

2. TF-IDF Calculation- To calculate the term frequency of the keyword, we focus only on the keyword input by the user rather than calculating the frequency of each word. This is done by using the stop word removal but in reverse i.e, we consider the keywords as the stop words and we focus on the stop words and ignore everything else while calculating the TF-IDF for our user keywords. We have manually assigned different weights to the terms occurring in title, abstract and author keywords based on the importance of the part of the document. We have given the highest importance to the term occurring inside the title followed by the term in author keyword which is then followed by terms in abstract.

doc = title + abstract + author keywords

TF-IDF(document) = TF-IDF(title)*(alpha) + TF-IDF(author keywords)*(theta) + TF-IDF(abstract)*(1-alpha-theta)

3. Proposed Algorithm

For i in range(0,len(D s)) do

- for j in range(0,len(D t [i]) do
- for k in range(0,len(D u)) do
- if D u [k] equals D t [j] then
- D s [i]←update t (D s [i])
- for j in range(0,len(D a [i]) do
- for k in range(0,len(D u) do
- if D u [k] equals D a [j] then
- D s [i]← update a (D s [i])
- for j in range(0,len(D p [i]) do
- for k in range(0,len(D u) do
- if D u [k] equals D p [j] then
- D s [i]← update p (D s [i])
- D s [i]←D s [i]+C[i]
- if D s [i] ≥ W min then
- Output←append (D t [i])

4. Working

The complete algorithm is divided into two parts i.e one which is used to produce recommendations for a new user and other one to bring diversity in the results for every time a user logs into the system.

Step (i) is the loop that will run for all the document weights from 0 to size of the dataset.

In steps (ii) to (iv), the matching of keywords is done between user profile and title keywords.

In steps(vi) to (viii), the matching is done between user profile and abstract keywords.

In steps(x) to (xii), the matching is done between keywords from user profile and author paper keywords. In steps (v),

(ix), (xiii) the weights have been updated according to the respective update functions as described in data structure section above. To bring the diversity in the results, the top results that come as recommendations from first part are used to extract the authors along with author keywords and we find papers from same authors having similar keywords. These papers are then recommended to the user on basis of high citation count for creating a diverse environment.

5. Implementation

We have tried to implement the above mentioned algorithm in the form of a chat bot where the bot will ask the user for their area of research and treat it as a keyword for the user profile. The bot also provides the option to request for only abstract of a paper and also for the pdf link of a paper. The complete implementation has been carried out in python and the use of libraries such as tensor flow, natural language processing etc has been made.

In near future, we aim to improve the chat bot to work in a broader range of topics and keywords rather than a few limited keywords in the existing system. The implementation has been performed using python 3 and we have used the Natural Language Processing library in python to process the data set and extract the required information from the data set. Only the important keywords are extracted and hence used for matching purpose. The user needs to enter a research area as the keyword and based upon that keyword the system will return the top 10 results to the user.

Now based upon the choice of user, they can either chose to view next 10 results, or they can request to view the abstract and pdf link for the article they like out of the show results. The top 10 results shown are the titles of the articles with various different options such as view abstract, pdf link, or view next 10 or even go back and change the research area. The main criteria used for the evaluation of top articles is based upon the keyword frequency in- side the title, abstract and author keywords with importance of title being highest and abstract being the lowest.

5. Results

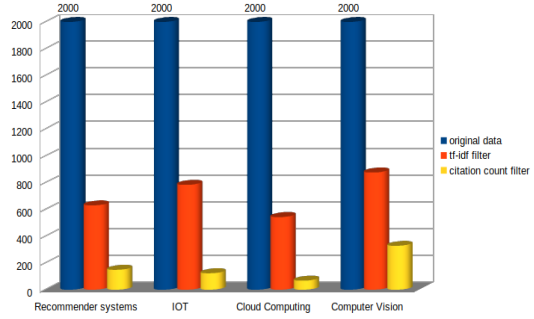

Figure 1 A comparison between number of results obtained without and with filtration.

The above figure denotes the number of final results received after applying the two types of filters in two different colors. The initial data size was 2000 which was reduced largely after the tf-idf filter and the citation count number was added.

Table 1 Tabular Representation of data

Research Field	Original Data	tf-idf Filter	Citation Count Filter
Recommender Systems	2000	630	149
IOT	2000	784	123
Cloud Computing	2000	5543	68
Computer Vision	2000	876	328

IV. CONCLUSION

The method studied above makes the use of bibliometrics from the articles and uses them to find the tf-idf score calculation to find the similarities between various articles and user interests. This way, the user keywords can be used to extract a particular set of fields in which the user is interested. With the help of the method used and explained above, it can be said that the goal has been achieved to reduce the cold start problem as we attain some information about the user and their interest of research area. The user enters the keyword for the area they are interested in and hence provide the system with some information about what they like.

This keyword is then used to find the tf-idf score match within the bibliometrics taken under consideration i.e. the title, abstract and author keywords. This eliminates the chances of the cold start problem as it provides at-least some information regarding what the user likes. This helps to generate the recommendations based upon what the user is interested to view and hence reduce the cold start problem. The data set used in this system is from only a single domain and is limited to a few research areas and specific keywords. The implementation has been carried out in form of a chat bot and this bot has features such as options to provide the user with the abstract as well as the link to the pdf of the article if the user desires to view it.

Future Scope

In near future, we tend to increase the data set to a wider domain and sources other than the ones being used currently. Also, system is limited to a few selected keywords only which will be expanded to a broader category in the near future. The bot implemented at current stage cannot keep the context of the conversation and hence it cannot answer a question related to the previous query. This maintenance of context can be added in the future. The system also lacks support for documents that link between IT and non-IT departments where the

techniques from one department are used in the other one. This can also be implemented in the near future.

REFERENCES

[1]. [Alepidou et al., 2011] Alepidou, Z. I., Vavliakis, K. N., and Mitkas, P. A. (2011). A semantic tag recommendation framework for collaborative tagging systems. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Inter- national Conference on Social Computing, pages 633–636.

[2]. [Anand, 2014] Anand, D. (2014). Evaluating folksonomy information sources for genre predic-tion. In 2014 IEEE International Advance Computing Conference (IACC), pages 887–892.

[3]. [Bahulikar, 2017] Bahulikar, S. (2017). Analyzing recommender systems and applying a loca- tion based approach using tagging. In 2017 2nd International Conference for Convergence in Technology (I2CT), pages 198–202.

[4]. [Dhanda and Verma, 2016] Dhanda, M. and Verma, V. (2016). Recommender system for aca-demic literature with incremental dataset. Procedia Computer Science, 89:483 – 491. Twelfth International Conference on Communication Networks, ICCN 2016, August 19 21, 2016, Ban-galore, India Twelfth International Conference on Data Mining and Warehousing, ICDMW 2016, August 19-21, 2016, Bangalore, India Twelfth International Conference on Image and Signal Processing, ICISP 2016, August 19-21, 2016, Bangalore, India.

[5]. [Hang and Meifang, 2011] Hang, C. and Meifang, Z. (2011). Improve tagging recommender system based on tags semantic similarity. In 2011 IEEE 3rd International Conference on Communication Software and Networks, pages 94–98.

[6]. [Kaur and Dhindsa, 2016] Kaur, S. and Dhindsa, K. S. (2016). Comparative study of citation and reference management tools: Mendeley, zotero and readcube. In 2016 International Conference on ICT in Business Industry Government (ICTBIG), pages 1–5.

[7]. Kohi et al., 2011] Kohi, A., Ebrahimi, S. J., and Jalali, M. (2011). Improving the accuracy and efficiency of tag recommendation system by applying hybrid methods. In 2011 1st International eConference on Computer and Knowledge Engineering (ICCKE), pages 242– 248.

[8]. [Kosala et al., 2017] Kosala, R., Ellen, and Saputra, M. D. (2017). Folksonomy-based image recommender for art commissioning. In 2017 International Conference on Applied Computer and Communication Technologies (ComCom), pages 1–5.

[9]. [Kumar, 2015] Kumar, A. (2015). Heuristic approach for recommending research articles using collaborative tagging. In International Journal of Innovations in Engineering and Technology (IJIET).

[10]. [ollagnier et al., 2018] ollagnier, A., Fournier, S., and Bellot, P. (2018). BIBLME RecSys: Harnessing Bibliometric Measures for a Scholarly Paper Recommender System. In BIR 2018 Workshop on Bibliometric-enhanced Information Retrieval, Grenoble, France.

[11]. [Sen and Riedl, 2011] Sen, S. and Riedl, J. (2011). Folksonomy formation. Computer, 44(5):97– 101.

[12]. [Tewari et al., 2016] Tewari, A. S., Yadav, N., and Barman, A. G. (2016). Efficient tag based personalised collaborative movie reccommendation system. In 2016 2nd International Con-ference on Contemporary Computing and Informatics (IC3I), pages 95–98.

[13]. [Wijonarko et al., 2017] Wijonarko, A., Nurjanah, D., and Kusumo, D. S. (2017). Hybrid recommender system using random walk with restart for social tagging system. In 2017 International Conference on Data and Software Engineering (ICoDSE), pages 1–6.

[14]. [Yoo and Suh, 2010] Yoo, D. and Suh, Y. (2010). User-categorized tags to build a structured folksonomy. In 2010 Second International Conference on Communication Software and Net-works, pages 160–164.

[15]. [Zhou, 2012] Zhou, Y. (2012). Extracting top hot hits in biomedicine from pubmed. In 2012 Fourth International Symposium on Information Science and Engineering, pages 266–269.