Hybrid News Recommendation System using TF-IDF and Similarity Weight Index

C.P. Patidar, Yogesh Katara, Meena Sharma

Abstract*: As the usage of internet is increasing, we are getting more dependent on it in our daily life. The Internet plays an essential role to simplify our tight schedules. In such tough lives, it is very important to stay aware of current affairs. Now for different people coming from different backgrounds and professions, the preferences are different too. Here come Data mining techniques in the picture, which gives us "Recommender system" as the output, capable of delivering more relevant and worthy outcomes. Newspapers are the basic obligation asked by almost every person to stay updated and aware of the world. But as we observe that nowadays, various solutions are been developed to convert paper news system to digital news and raise the bar of the quick news. And that's how News Recommender systems are have made an important place in our fast running lives.This research paper has investigated the News Recommendation solution right from its core, including the importance, performance, and improvement suggestions. This paper talks about enhancing the performance of states solution by using modified Term Frequency-Inverse Document Frequency (TF-IDF) algorithms. Proposed solution advocates the usage of JAVA technology which reflects fruitful results in the final graphs of accuracy, precision, and F-score. Here, BBC dataset has been used for comparison study purposes.

***Keywords*: Associative Calculus, BBC Dataset, News Recommendation, TF-IDF.**

I. INTRODUCTION

The process of data extraction is called Data Mining. Later, this technique is used by different algorithms as their raw feed to result out better outcomes with précised and more relevant content. Data mining is classified into two forms: Descriptive Analysis & Predictive Analysis. Both techniques are capable of fetching desired results with great efficiency.

Recommendation system provides the user with the content they prefer to learn about. For this purpose, this system processes both separate and specialized set of data. From the past a few years, we have observed that personalization is taken on a whole new level and now the user is assisted with more precise and overload data according to their preferences.This works also advocates the incorporation of Data mining & Recommendation system, which can help in fetching the processed data purely extracted from the user's preference and today's trend cluster. This solution not only provides a better system but also better results with popularity factors and trend results.

Revised Manuscript Received on October 05, 2020.

***** Correspondence Author

C. P. Patidar, Assistant Professor, Department of Information Technology, IET DAVV, Indore, India. Email: cpatidar@ietdavv.edu.in

Yogesh Katara*, PG Student, Department of Information Technology, IET DAVV, Indore, India. Email: yogeshnayak18@gmail.com

Dr. Meena Sharma, Professor, Department of Computer Science, IET DAVV, Indore, India. Email: msharma@ietdavv.edu.in

News Recommendation system comprises of relevant news, posts and suggestions purely based on the user interest. This system can offer news, supported news, quality, and visits. The other factors used by News Recommendation system are news ranking, area, impact, priority etc[4]. The most basic expectation from a recommendation system is to cater to multiple results based on the similarity factor. For example, if a user is looking for news related to sports for a particular team, the system must recommend all respective news belonging to that word along with all co-relevant news based no currently searched topic. This system aims all the relevant news based on the search and user demand word, along with the news which was not actually asked but stays relevant to the topic. You can consider YouTube for a perfect example, in their news section, they also recommend the extra news in the below section about the same topic you searched. Collaborative filtering, Content mining, Classification of Association rule are the various solutions which can be applied under the label of Recommendation system. The system has considered TF-IDF as the foundation algorithm for content mining and customized algorithms are used for document mapping purposes.

In this research paper, a unique hybrid approach is been introduced by using customized algorithms for document mapping and TF-IDF algorithm. The coming next section talks about the basic study of the previous solution and associated problem. Then, we have discussed the Problem, Proposed Solution and also comparison graphs, analyzing various factors has been provided. This paper wraps itself with conclusion and stating the references used.

II. RELATED WORK

A Recommendation System proposed by Maria Bielikova & Michal Kompan et al. [2] claims to extract relevant news precisely based on user preferences. They used Slovak News Portal to explain their solution for the news recommendation system based on content mining. Their proposed solution divided the news information into two different sections named as Article and User Activity, to generate the personalized recommendations. The main principle of this solution is based on the article similarity algorithm which includes keywords, category, content, names/places etc. In their first step, they preprocess the News articles. Then as the second step, the recommendations are made based on the ratio of recommended and visited articles. And as the final third step, recommendations are made based on the ratio of recommended and non-visited articles.

Although, this system may recommend a good result, they have not used some key matrices like relevancy and popularity of news, which surely lower downs the standards. Their proposed solution is represented in Fig. 1.

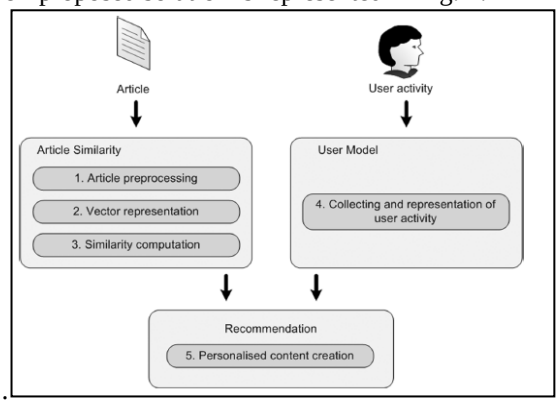

Fig. 1. News Recommendation System[2]

Factors included in implementation of the above mentioned solution, are;

- Title
- TF of Title words in the article content
- Keywords finding
- Names and Places extraction
- CLI Coleman-Liau Index
- Category extraction

Adomavicius et al. [3] address that recommender systems are becoming increasingly important to individual users and businesses for providing personalized recommendations. They investigate that most of the researchers have only focused on recommendation accuracy, other important aspects of recommendation quality, such as the diversity of recommendations, have often been overlooked. It also suggests that the recommendation system are highly important in the current world scenario as data on browsers is very huge. Individuals, as well as business, need a class level of recommenders. Investigations observed that most of the researchers have focused on the accuracy of recommendation, quality and diversity are mostly ignored. In this paper, the recommendation is given based on accuracy as well as based on item ranking techniques that can generate more fine results. Deeper studies of conventional algorithms suggest that TF-IDF is only capable of telling how many times a certain word appears in the given document. Bahram amini, et al. [9] focuses on user search in a recommendation system. User profile plays a major role infiltration techniques as user profile signifies what one can search. User logs are a wide collection of data hence searches should be specific. This study gives a brief overview of a recommender system.

III. PROBLEM DOMAIN

Heavy load of Data Mining techniques has been observed in the complete study, in the process of enhancing the Data generation. It has become more complex in finding a very specific and accurate level of result. Back then, Content mining & TF-IDF algorithms are been simultaneously used for relation finding and similarity extraction purposes. Studies have found that such algorithms are limited to work on the principle of word similarity only, while a relationship with documents and involvement is completely ignored. Every person has a different mindset and distinct reading bias, in such cases, the results must get varied according to the user preferences and demands. At the same time, impact and quality of data play a necessary role in user search too. A News Portal is divided and sub-divided in various sections entirely based on the nature of the news content and similarity of news articles. All sections have a different nature yet similar importance. Content-based News Recommendation system is entirely dependent on the author, which lacks both accuracy and relevancy. This work revolves around article similarity rule and assessed on a non-English database. And hence, for this reason, English database evaluation with document mapping is anticipated. The studies till now have suggested to us that we need a solution which demands a strong need to revise content mining approach to document-based now, from word-based. At the same time, the proposed solution must be capable of evaluating similarity and accuracy for English database.

IV. SOLUTION DOMAIN

A major part of our daily lives is secured by News nowadays. News is said to have its full form as North East West South, that denotes that news is making you aware of all the directions of the world. On different platforms, News gets catered such as newspaper, television, magazine etc. News makes us stay updated about the country and world we are living in. Selection of news can be subjective and entirely dependent on the choice of the user, people are free to read the news about the topics they like or find relevant to their interest. The news recommendation system is used to have the desired information while searching. Different news content may have different news category. Sometime, the news category may be known before recommendation but sometimes no one knows about news category. We have to use a learning approach to identify the category of news and recommend them according to relevancy factor. A Hybrid Solution using Machine Learning based Naïve Bayes Classification technique along with TF-IDF algorithm has been proposed to make a common and most relevant recommendation. A block diagram to explore this solution is shown in Fig 2. This research paper will attempt to integrate the concept of machine learning-based naïve Bayes algorithm to categorise the news based on previous training and classify them according to their nature. Here, training will be done to specify the nature of different news article to identify their category. It will use unknown news collection and classifier will categories them into a different category. Afterwards, the user will provide their liking category and TF-TDF will perform to identify most closed news based on user input keywords. Core parts of information mining technology are under the technology blade for further improvements and development for several years such as computing, statistics and machine learning. A recommendation system comprising of collaborative filtering and TF-IDF algorithms can turn out to be a great option too, but this has its own set of the scope of improvement. A separate association rule or similarity matching algorithm shall be added to improve the performance of the system.

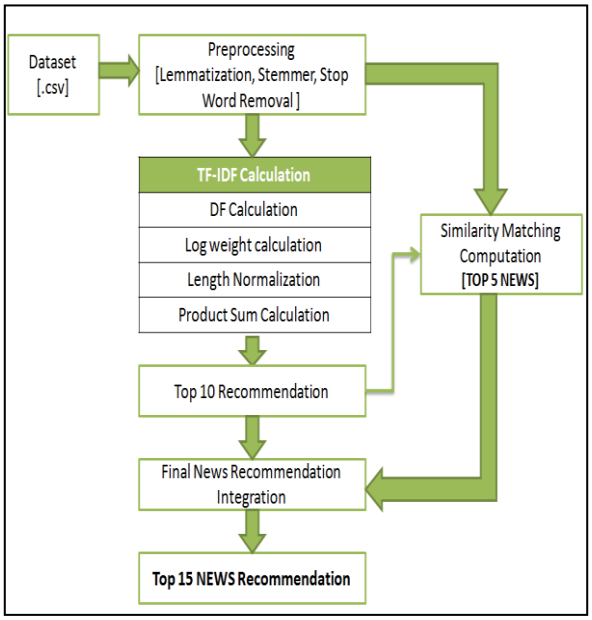

Fig. 2.Proposed News Recommendation System

A pool of data and half-baked knowledge of search tools results in wasteful data retrieval and failure in desired information extraction. User preference-based précised practice, specialized and separate set of information are offered by the Recommendation system. In past years, internet users are saved from the data overload problem, all thanks to the web personalization. The complete study proves that web personalization can do wonders in the field of News Recommendation, as this is effective in hiking the quality of content mining and the system then recommends more useful and relevant bracket of information.

The three modules in which the proposed solution is divided, are as follow:

1. Implementation of TF-IDF algorithm.
2. Integration of Similarity Matching & Computation Approach with TF-IDF.
3. Performance evaluation and NEWS Recommendation.

The proposed algorithm of this work has been shown in Fig. 3,4.

Pseudocode 1: TF-IDF Algorithm

1. for(i=0, i<numnberOfUniqueWords, i++) //finds number of unique words in documnets
2. for(j=0, j<numberOfDocuments, j++) //loop for the total documents
3. tfidf = f_{ij}_ log(numberOfDocuments = n_i) //calculation of Tf-Idf
4. for(s=0, s<numnberOfUniqueWords, s++)
5. f_{ij}_Temp = number_of_occurrences_of_word_S _in _the_document_J
6. tfidf_Temp=f_{ij}_Temp* log(numberOfDocuments / df_i)
7. sum_Tfidf += (tfidf_Temp)^2
8. end
9. A[i,j] = tfidf/sum_Tfidf
10. end
11. end

Fig. 3. Pseudocode of TF-IDF Algorithm

Pseudocode 2: Similarity Matching

1. D1={d1,d2,d3…..dn}//Document list
2. D2={d1,d2,d3…..dn}// Another Copy of D1
3. WDn = {wd1,wd2,wd3…wdn} //WDi = wordlist of Di
4. Ac={wd1||wd1,wd2,wd3…wdn}+{wd2||wd1,wd2,wd3…wdn}…WDn

Fig. 4.Pseudocode of Similarity Matching

The proposed solution integrates both Pseudocode 1 and Pseudocode 2 into a single module and suggests a hybrid approach for news recommendation. This discussed work is been implemented for BBC database of English language, with more than 2100 transactions under test. In the first step, the IR approach is been implemented using TF-IDF algorithm. After the successful execution of step one, for similarity matching purposes, a document mapping algorithm is implemented. Here, every word from one document is matched with other documents. Let us take 2 different documents, for example, A & B. Let's consider A has a sentence with words I LIKE INDORE and document B has a sentence with words I LIKE BHOPAL. So, both documents will get compared for each word and both I & LIKE will come out as common in both the documents. The same way, the comparison is expected in the proposed solution. The complete solution will give us an integrated module of TF-IDF and similarity matching algorithm.

V. RESULTS AND OBSERVATIONS

JAVA technology is used in the evaluation and implementation of this work. To evaluate the performance of this proposed solution, it is tested on parameters like Precision, Accuracy & Final Score. Below are the different categories of news database, used to evaluate the solution:

1. Politics
2. Business
3. Sports
4. Technology
5. Entertainment

A set of 5 words from each category are taken to explore the accuracy and relevancy of the proposed system.
The complete input data is presented in Table I.

Table- I: Data Input

Category	Word
Politics	Narendra Modi, Rahul Gandhi
Business	Corporate, IT industry
Sports	Wankhede stadium
Technology	Security, Online Classes
Entertainment	Award, Bollywood

The above-shown Table has all the input source, every repetition demands the final score of relevant categories must be higher than the previous solution. To fulfill this demand, a competitive result analysis have been performed, where outcomes has been matched with the desired category.

Table- II: Comparison of Previous Algorithm

BBC-Dataset	TF-IDF	Previous Work	Proposed
Best Accuracy	0.85	0.589	0.956

Fig. 5, 6, 7 shows the comparative study of all the evaluated results.

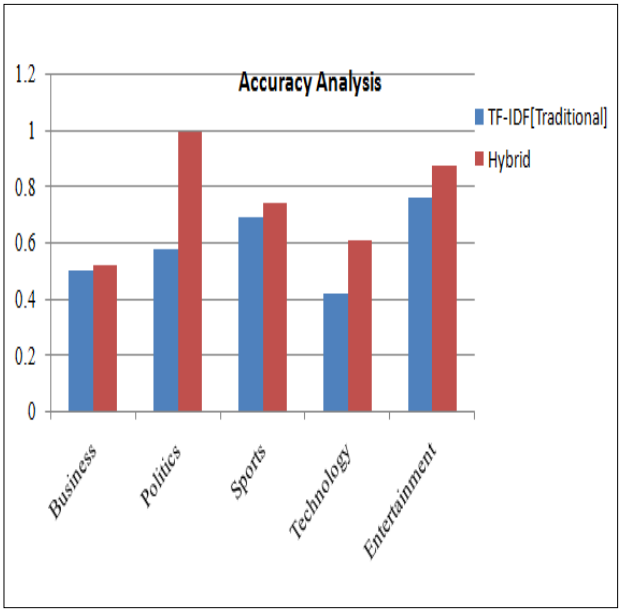

Fig. 5. Accuracy Comparison between traditional and proposed

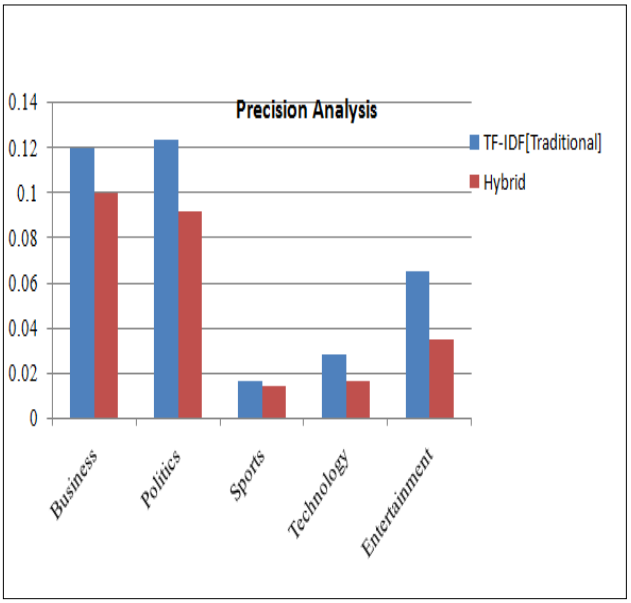

Fig. 6. Precision Comparison between traditional and proposed

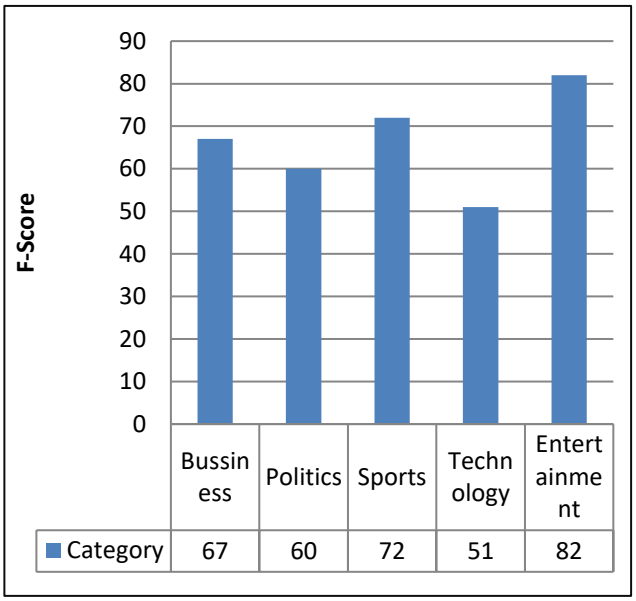

Fig. 7. Final Score of Hybrid Algorithm

VI. CONCLUSION

Because Recommendation system has surely made some great progress in the past few years and plenty of new techniques and algorithms are introduced to boost the reference quality. We have observed that every new technique is developed to enhance the recommendation accuracy but losing the side of recommendation diversity. The planned system will not only recommend the news content according to the user preference but also refine the news articles based on impact and priority.

This proposed solution is all set to be tested and part in expansion with news datasets like TOI or The Hindu. The final score in the graphs and tables confirm the improvement in the accuracy.

REFERENCES

1. Neeraj Raheja, V.K.Katiyar," International Journal of Computer Science Issues" Vol. 11, pp-2, 2014.
2. Kompan M., Bieliková M. (2010) Content-Based News Recommendation. In: Buccafurri F., Semeraro G. (eds) E-Commerce and Web Technologies. EC-Web 2010. Lecture Notes in Business Information Processing, vol 61. Springer, Berlin, Heidelberg.
3. Adomavicius, G, & Kwon, Y. O. (2012). Improving aggregate recommendation diversity using ranking-based techniques. IEEE Transactions on Knowledge and Data Engineering, 24(5), 896-911.
4. MinsukKahng, Sangkeun Lee, Sang-goo Lee, Ranking in Context-Aware Recommender Systems,pp-65-66, 2011.
5. Y. Ma, S. Ji, Y. Liang, J. Zhao and Y. Cui, "A Hybrid Recommendation List Aggregation Algorithm for Group Recommendation," 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Singapore, 2015, pp. 405-408.J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.
6. Ch.Nagini, M.Srinivasa Rao, Dr. R.V.Krishnaiah, International Journal of Engineering Research & Technology, Vol. 2,pp-701-704,2013.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," *IEEE Transl. J. Magn.Jpn.*, vol. 2, Aug. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301].

7. Barskar, N. and Patidar, C.P., 2016. A survey on cross browser inconsistencies in web application. International Journal of Computer Applications, 137(4), pp.37-41.
8. Gediminas Adomavicius, Young, Kwon Improving Recommendation Diversity Using Ranking-Based Techniques, pp-1-33.
9. Bahram amini, roliana ibrahim, mohd shahizan othman, "Discovering the impact of knowledge in recommender systems: a comparative study", International Journal of Computer Science & Engineering Survey, vol 2,pp-3,2011.
10. Patidar CP, Sharma M. An Automated Approach for Cross-Browser Inconsistency (XBI) Detection. InProceedings of the 9th Annual ACM India Conference 2016 Oct 21 (pp. 141-145). ACM.

AUTHORS PROFILE

Chandra Prakash Patidar received his Bachelor's. degree in Information Technology and M.E. degree in Computer Engineering. He is an Assistant Professor of Information Technology at the Devi Ahilya University, Indore, India. His research interests are in Cross Browser Testing, GPGPU Computing, CUDA Programming, Multithreaded Architecture, Compiler and Memory Architecture of Computers.

Yogesh Katara received his Bachelor's degree in the department of Computer science and Engineering from IES IPS Academy, RGPV University. Currently He is Pursuing his Master's degree in DAVV University, Madhya Pradesh. His specializations include network security, Machine Learning, Database management systems, Data mining and IOT.

Dr. Meena Sharma received her Bachelor's degree in Computer Engineering and M. Tech degree in Computer Science in 1992 and 2004 respectively. She received the Ph. D. Degree in Computer Engineering in 2012.She is a Professor of Computer Engineering at the Devi Ahilya University, Indore, India. Her research interests are in Software Engineering, Software Quality Matrices and Object Oriented Modelling and Design.