ISSN No:-2456-2165

# Hybrid News Recommendation System using TF-IDF and Machine Learning Approach

C.P. Patidar, Dr. Meena Sharma, Yogesh Katara
Department of Information Technology,
IET DAVV, Indore M.P., India

**Abstract:- A newspaper divided into various sections like a city, sports, editorial, international, national, entertainment etc. All the above sections have equal importance and different user followers for different sections. Sometimes there may be a possibility that they may consist of relevant information but in different sections and different newspapers. News Recommendation System can overcome this problem and suggest relevant news according to user preference and popularity factor. This research paper investigates the need for news recommendation using a machine learning approach to make it more efficient and better. Hybrid Approach can help to recommend news to users based on Supervised Machine Learning and Term Frequency-Inverse Document Frequency (TF-IDF).**

***Keywords:-*** *Machine Learning, Naïve Bayes, News Recommen- Dation, TF-IDF.*

## I. INTRODUCTION

Tosearch for desired information user faces difficulty because of getting irrelevant information. This issue arises due to the insufficient knowledge of search tools and availability of a large amount of data. Extraction of desired information becomes difficult in this case. The recommendation system is beneficial in this case, which offers a relevant set of information. The examination analyzes, a broad application or device that includes client inclination or self-gathered information for predicting client's need and investigates the best probability of importance among data which is known as recommendation System, or it tends to be express that recommendation system is the device that gives pre-determined information-based data. Recommendation system may valuable in different fields, for example, news, shopping, item search and so on.

The recommendation system utilizes various advancements. Recommendation system is classified as:

1. Content-based recommender system works on user preference and content exists at the data source. It compares and extracts the information from web pages and data sources and matches with user preference. It also uses popularity calculations and frequent uses to find the most used and most demanding content. It uses this concept to evaluate and sort content according to demand and popularity. Generally, it observes the description associated with items or existing content and compares with user preference.
2. Collaborative separating frameworks prescribe things dependent on comparability criterion among clients or potentially things. The things prescribed to a client are those favoured by comparable clients. This kind of recommendation can utilize the preparation on likeness search and bunching. Nonetheless, these innovations without anyone else's input are not adequate, and some new calculations that have demonstrated compelling for recommendation system.
3. Knowledge frameworks prescribe proposals or arrangement by producing physically or naturally various ends and choice standards. It stresses on express field information about the necessities and client inclination. On the other hand, physically created choice principles or made inferences might be one-sided and not appropriate for customized frameworks. This framework related to various downsides, for example, bottleneck issue during information handling and acquire issue during client profile creation and connecting with existing data. A programmed information-based framework is prescribed where the contribution of information might be emotional and can fluctuate as indicated by prerequisite.
4. A demographic recommender system gives suggestion dependent on a client's statistic profile, which includes the client's statistic information, for example, sex, age, date of birth, instruction and other individual highlights. This methodology classifies clients into bunches dependent on their statistic attributes and suggests protests as needs are. All the more definitely, it accept clients in a similar classification have a similar taste or inclinations. Proposals are given for new clients by first recognizing the classification client have a place with and afterwards by finding inclinations of different clients in a similar class.

## II. RELATED WORK

Different recommendation system techniques and their pitfalls are mentioned in table1.

TABLE 1
LITERATURE REVIEW

| S. No | Title | Tools/Techniques | Concept | Advantage | Disadvantage |
|---|---|---|---|---|---|
| 1. | Cold Start Recommenda-tion Based on Attribute-Fused Singular Value Decomposition [1]. | KNN, Collaborative Filtering, Matrix Factorization. | Combines with the attribute information of the item with the historical rating matrix to predict the potential preferences of the user. | A significant improvement in recommendation accuracy with solving cold start problem of new items. | Cold start problem. |
| 2. | A Recommendatio n Model Based on Content and Social Network [2]. | Recommendation model based on content and social network (RMBCS). | Proposed a new distance to calculate the text similarity between long text and short text. After that nearest neighbour group is found from user's social network. Then, recommend the texts. | Improves the accuracy of the recommendation, reduces the cost and cost of training, also enhances the novelty of the recommendation. | It does not optimize the similarity of text and improve recommended performance. |
| 3. | A New Collaborative Filtering Recommendatio-n Algorithm Based on Dimensionality Reduction and Clustering Techniques [3]. | K-means algorithm and Singular Value Decomposition (SVD). | It proposes and evaluates an effective two stage recommender system that can generate accurate and highly efficient recommend-ations. | Significantly improved the performance of recommendations and remained the lowest values in the RMSE curve in the whole neighbours range. | Scalability is not achieved. |

Y. Ma, et al. [4] described that in group-oriented recommendation field, the design of a commonly acceptable recommendation list is a tough task. Traditional group recommendation algorithms those are used in the recommendation are often realize group recommendation list aggregation according to the item ranking or their item score of group members' recommendation lists. The factors that are considered in these algorithm is relatively one-sided.

L. Zonglei, et al. [5] demonstrated a new method to forecast flight delays. This new method is based on the content-based recommendation system. In the forecast model, the events such as flight delays and airportstn that have been mapped to users and items, respectively, which are the concepts in the recommendation system. According to the propagation of delay, this method alerts the target airport by monitoring the status of the related airports. The observed status is compared with the historical data to predict the seriousness of delay. Since the airborne hours between every two airports are usually more than an hour, this method could give the alarm at least 1 hour ahead. Besides, the above factors the method requires minimum online calculation, and therefore it guarantees that the delay forecast can be delivered in a quick and timely manner.

Bahram amini, et al. [7] 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. Data from different sources which are searched is considered in this work. Personalization system is classified in several ways some are utility functions or call modelling. These work further describe a hybrid approach which combines content-based, collaborative based and knowledge-based approach. The knowledge-based system generates recommendation with the help of decision rules. All the specifications of users are analyzed then knowledge-intensive rules that are generated based on users choice having similarities. Traditional content-based recommendation uses the data on web pages and ratings of it which user browsed. The comparison of user profile and data on the web pages are done. The traditional scheme follows the pattern in which it strongly believes that new choices are highly influenced by the past.

Adomavicius, et al. [8] 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.

## III. PROBLEM DOMAİN AND OBJECTİVES

News reading is one of the most common activity of daily life. The developing web world and rushed timetable of day-by-day life make such a great amount of trouble for web users to discover related news. This circumstance turns out to be more regrettable when client attempt to inquire data and get immaterial news content. Insufficient learning of pursuit machine and an extensive measure of information gives poor execution to recover or separate news content. Suggestion frameworks offer scholarly practice in view of client inclination. Proposal frameworks offer a discrete and specific arrangement of data. As of late, Web personalization for news has gotten much regard for help Internet clients with the issue of data over-burden.

Following points are expected from proposed research work:

- To load and clean data of BBC news dataset and load for lemmatization and filtering.
- To implement the Naïve Byes classification algorithm to classify the data into multiple categories.
- To implement TF-IDF algorithm and recommend news articles accordingly.
- To recommend news articles based on classification and TF-IDF algorithm along with estimate results based on accuracy, precision and f-score.

## IV. PROPOSED WORK

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 depicted in Fig 1.

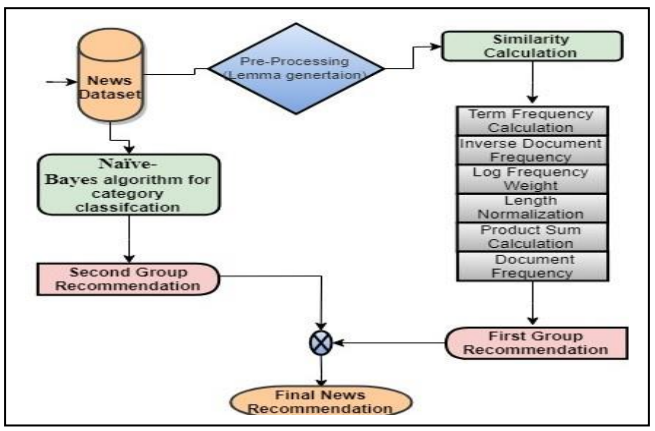


Fig 1:- Flow of the proposed recommendation system

The complete work has been classified into four modules which are following:

**Module 1: Dataset**

BBC Dataset has been recommended to consider as the input source for news recommendation.

**Module 2: Classification using Machine Learning**

Machine learning is used to first learn the concept and then apply intelligence to take a decision. This module will help to learn about the thought behind categorization of news and relevance to categorize unlabeled news into categories. Here, the Naïve Bayes classification algorithm has been used to classify the data into multiple categories. Initially, a training step will be used to provide learning to data then work will be classified according to learn thought and during the testing module. Then after user desired category news will be forwarded to the next module for the top recommendation.

**Module 3: TF-IDF Algorithm**

TF-IDF is an information retrieval(IR) algorithm based on the occurrence of keywords in the whole dataset as well as particular documents. A detailed description of the DF calculation is cited below;

➢ **Calculation of Document Frequency**

- The term frequency $tf_t$, d of term t in the document d can be defined as the number of times that t occurs in d.
- A document with the tf = 10 occurrences of the term is more relevant than a document with the
- tf = 1 occurrence of the term.
- Relevance does not increase proportionally with term frequency.
- The document frequency can be defined as the number of documents in a collection that the term occurs in.
- $df_t$ is the document frequency can be defined as the number of documents that t occurs in.
- $df_t$ can be defined as an inverse measure of the informativeness of term t.

➢ **Calculation of Log frequency weighting**

Following steps are executed to calculate log frequency weighting:

- Equation (1) shows the log frequency weight of term t in d.

$$W_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d} & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

- Score for a document-query pair can be given as a sum over terms t in both *q*and *d*:

Tf-matching-score(*q*, *d*) can be calculated as:

$_{t \in q \cap d}$ $(1 + \log tf_{t,d})$

- The score is said to be 0 if none of the query terms are present in the document.

➢ **Calculation of IDF [Inverse Document Frequency]**

Equation (2) shows the IDF weight of term t.

$$idf_t = \log_{10} \frac{N}{df_t} \quad (2)$$

**Important Points:**

- Wher *N* is the number of documents in the collection.
- $idf_t$ is a measure of the informativeness preset in the document of the term.

➢ **Calculation of TF-IDF weighting**

Equation (3) shows the tf-tdf weight of a term is the product of its tf weight and its idf weight.

$$W_{t,d} = (1 + \log_{10} tf_{t,d}).\log_{10} \frac{N}{df_t} \quad (3)$$

Outcome of this module generates the product sum for every document which will help to evaluate top ten News recommendation.

**Module 4: Similarity Matching & Recommendation**

This module will take user choice in terms of keywords and threshold value to decide the cutoff for the recommendation. A final bunch of news will be recommended as a final output.

## V. CONCLUSİON

This research work addresses the need of modern news recommendation system based on user choice. This research work identifies that machine learning approach could help to classify the news into multiple categories and TF-IDF could help to find the similarity factor and decide most relevant news. A model of the proposed solution is also developed and define inside proposed work. This work will be implemented using java technology and it will be evaluated based on precision, recall and f-score along with computation time to measure computation performance.

### ACKNOWLEDGMENT

This paper and the research behind it would not have been possible without the exceptional support of my supervisor, Dr. Meena Sharma and Mr. C.P.Patidar. His enthusiasm, knowledge and attention to detail have been an inspiration and kept my work on track.

### REFERENCES

[1]. X. Guo, S. Yin, Y. Zhang, W. Li and Q. He, "Cold Start Recommendation Based on Attribute-Fused Singular Value Decomposition," in IEEE Access, vol. 7, pp. 11349-11359, 2019.

[2]. H. Xue and D. Zhang, "A Recommendation Model Based on Content and Social Network," 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2019, pp. 477-481.

[3]. H. Zarzour, Z. Al-Sharif, M. Al-Ayyoub and Y. Jararweh, "A new collaborative filtering recommendation algorithm

based on dimensionality reduction and clustering techniques," 2018 9th International Conference on Information and Communication Systems (ICICS), Irbid, 2018, pp. 102-106.

[4]. 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.

[5]. L. Zonglei, W. Jiandong and X. Tao, "A new method for flight delays forecast based on the recommendation system," 2009 ISECS International Colloquium on Computing, Communication, Control, and Management, Sanya, 2009, pp. 46-49.

[6]. 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.

[7]. 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.

[8]. 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.

[9]. 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.

[10]. Aizawa, A., 2003. An information-theoretic perspective of tf–idf measures. Information Processing & Management, 39(1), pp.45-65.

[11]. J. Liu, M. Tang, Z. Zheng, X. F. Liu, and S. Lyu, "Location aware and personalized collaborative filtering for web service recommendation," IEEE Transactions on Services Computing, vol. 9, no. 5, pp. 686-699, 2016.

[12]. J. Wang, A. P. De Vries, and M. J. Reinders, "Unifying user based and item-based collaborative filtering approaches by similarity fusion," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 501-508.

[13]. Patidar, C., Sharma, M. and Sharda, V., 2017. Detection of cross browser inconsistency by comparing extracted attributes. International Journal of Scientific Research in Computer Science and Engineering, 5(1), pp.1-6.

[14]. M. Aleksandrova, A. Brun, A. Boyer, and O. Chertov, "Identifying representative users in matrix factorization-based recommender systems: application to solving the content-less new item cold-start problem," Journal of Intelligent Information Systems, vol. 48, no. 2, pp. 365-397, 2017.

[15]. S. Deng, L. Huang, G. Xu, X. Wu, and Z. Wu, "On deep learning for trust-aware recommendations in social networks," IEEE transactions on neural networks and learning systems, vol. 28, no. 5, pp. 1164-1177, 2017.