International Journal for Rapid Research in Engineering Technology & Applied Science
Vol VI Issue II February 2020
ISSN (Online): 2455-4723

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

ABSTRACT

A News Paper have many parts or sections such as sports, entertainment, advertisements, national, international and local news. This all parts or sections of news paper have their own importance. Sometimes they have related information but in different section or different newspaper. To overcome from this problem they follow News Recommendation System. This research paper investigate the need of news recommendation using machine learning approach to make it more efficient and better, Hybrid Approach can help to recommend users based on Supervised and Unsupervised using Machine Learning and TF-IDF.

Keywords: TF-IDF, Machine Learning, News Recommendation

1. INTRODUCTION

User faces many difficulties because of getting irrelevant information when they search for suitable information. This problem occur because of insufficient knowledge of search tools and availability of large amount of data. In this case extraction of desired information becomes difficult and Recommendation system is beneficial which offers with a related set of information. The examination analyze, a broad application or device. This device or broad application 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. Recommendation system is important in different fields such as news, shopping, checking, item search and so on. Recommendation system also gives predetermined information based data. Recommendation system utilizes various development. Recommendation system is classified as follows:

1. Content-based system look at properties of the things suggested. For example, in the event, Netflix client has viewed various rancher motion picture. At that point grouped all pictures in the database as having the "cattle rustler" sort.

2. Collaborative separation frameworks determine things dependent on a comparative gauge between customers or potential things. Things offered to customers are preferred by comparable customers. Such a recommendation may use similarity search and bunch preparation. However, these innovations are not sufficient without anyone else's input, and there are some new calculations, which have demonstrated attractive performance for the recommendation system.

3. The structures of knowledge determine the propositions or arrangements by physically or naturally constructing different ends and choice standards. It emphasizes express sector information about requirements and customer orientations.

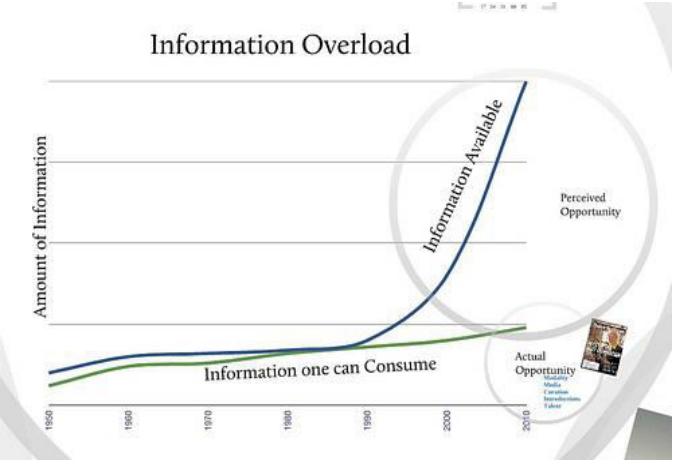
4. On the other hand, the principle of choice or infrared made physically may be unilateral and may not be suitable for customized frameworks. This framework related to various frameworks,

for example, the bottleneck problem during information profiles and the problem during client profile creation and linking with existing data. A programmed information-based framework is defined where the contribution of information can be emotional and fluctuate according to prerequisites.

5. A demographic recommendation system suggests relying on a client's statistical profile, which includes the customer's statistical information, for example, gender, age, date of birth and other personal highlights. This methodology classifies customers into clusters dependent on their statistical characteristics and suggests opposition as needed. All in all, it accepts customers in a similar assortment to have a similar taste or inclination. It is proposed for new clients that by first identifying, the classification client has the same class with finding the inclination of different clients and then having a place.

2. RELATED WORK

Gediminas Dominicus [3], he developed a recommendation system counting the ranking of news. They used the parameterized ranking approach to find articles that go beyond the limits. Recommended news based on ranking factor but relevance of news content is missing. A comparative study



Yuvkun Ma et al. [4] shows that the herd is documented in the proposal, usually the scheme of qualified suggestion lists is an extreme error. Conventional gathering suggestion calculations often accept the total proposal list according to the talk scores or talk scores of individuals gathering records. The elements considered in these calculations are moderately uneven. This paper advances another HAAB collection calculation for the list, which considers the state of things as the score of the proposal records of individuals. The test results show that the HAAB calculation can clearly beat the traditional assembly suggestion calculation, determining for different regular mixtures of different assemblies.

L. Zonglei et al. [5] demonstrated another strategy for estimating flight delays. This new strategy relies on a content-based suggestion framework. In the estimation model, opportunities "flight deferrals" and aerial terminals are mapped separately to customers and things considered in the suggestion framework. As the postponement signal suggests, this new strategy replaces the objective air terminal by inspecting the status of the associated air terminal. Since the air hours between each two air terminals are typically more than 60 minutes, this strategy may give a warning at any rate 1 hour in

advance. Additionally, the technique requires at least online computation, and along these lines ensures that the differential approximation can be expressed in a flamboyant and convenient man.

Bahram Amini et. al. [4] Center around customer discovery in recommendation systems. The customer profile assumes an important function in filtration processes because the customer profile means what one can see. The client log is a wide assortment of information, as a result of which the search must be clear. This exam gives a brief diagram of the recommender system. Information from various sources, which is seen, is considered in this work. Personalization frameworks are classified in a few different ways, some being utility or call performance. These works further illustrate crossover approaches that incorporate content-based, shared-based and information-based approaches. The methods and their nets of the diversity proposal framework are referred to.

3. PROBLEM DOMAIN

Reading the news is one of the most common activities in daily life. The timetable of the developing web world and day-to-day life causes so much trouble for web users to search for related news. This situation becomes more disconcerting when the customer tries to query the data and obtain immaterial news content. Inadequate learning of the stalking machine and extensive measurement of information gives poor performance to retrieve or separate news content. Suggestion frameworks offer scholarly practice in view of customer orientations. The proposal framework has discrete and data-specific arrangements. Of late, web personalization for news has gained a lot of respect for helping internet customers with the issue of data over-burden. The following points are expected from the proposed research work.

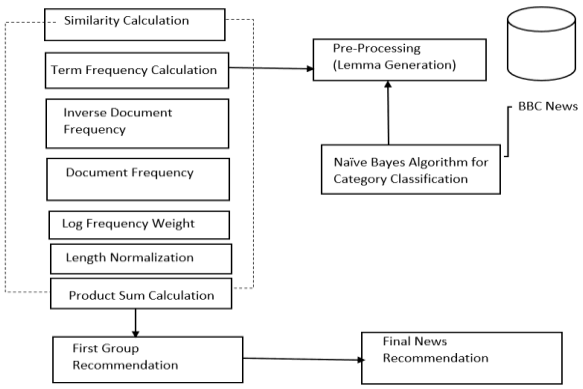
To load and clean data of the BBC dataset, and for lamination and filtering.

1. Analysis and research conclude that if web personalization is included in the news recommendation system then the quality of content mining will improve and more useful information will be recommended. This will result in taking the best available useful information.

4. PROPOSED WORK

The news recommendation system is used for the desired information while searching. Different news content may have different news categories. Sometimes, the news category can be known before the recommendation but sometimes no one knows about the news category. We have to use the learning approach to identify the category of news and recommend them according to the relevance factor. A hybrid solution using machine learning based Naive Bayes classification technique with TF-IDF algorithm has been proposed to be a common and most relevant recommendation.

A block diagram to find this solution is quoted below:



This research paper will attempt to integrate the concept of machine learning based web algorithm to classify news based on previous training and classify them according to their nature. Here, a training will be given to specify the nature of various news articles to identify their category. It will use anonymous news collections and the classifier will classify them into various categories. Later, the user will provide the category of their choice and TF-TDF will perform to identify most closed news based on user input keywords.

The complete work has been classified into four modules which are cited blow;

Module 1: Dataset

The BBC dataset is recommended to be considered as an input source for the news recommendation.

Module 2: Classification using Machine Learning

Machine learning is used to learn the concept first and then to apply intelligence to decision making. This module will help to know about the classification of news and the relevance of categorizing non-listed news into categories. Here, Naive Bayes Classification algorithm is used to classify the data into several categories.

Initially, a training step will be used to learn the data then the work will be classified according to the learning test and during the test module. The news of the user desired category will then be sent to the next module for top recommendation.

Module 3: TF-IDF Algorithm

TF-IDF is an IR algorithm based on the occurrence of keywords in entire datasets as well as in special documents. A detailed description about the DF calculation is quoted below;

- **Calculation of Document Frequency**

  1. Document d in period d of period frequency tft, d is defined as the number of times that t occurs.
  2. A document with tf = 10 frequencies of the word is more relevant than a document with the occurrence of the word tf = 1.
  3. Relevance does not increase proportionally with word frequency.
  4. Document frequency is the number of documents in a collection that are in the word.
  5. dft is the document frequency, the number of documents occurring in t.
  6. dft is an inverse measure in the drafting of term t.

- **Calculation of Log frequency weighting**

Following steps are executed to calculate log frequency weighting.

1. The log frequency weight of term T in D is defined as follows

$$w_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{tf}_{t,d} & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

2. $\mathrm{tf}_{t,d} \rightarrow w_{t,d}$:
   $0 \rightarrow 0$, $1 \rightarrow 1$, $2 \rightarrow 1.3$, $10 \rightarrow 2$, $1000 \rightarrow 4$, etc.
3. Score for a document-query pair: Sum of words in both q and d: tf-matching-score (q, d) = t∩q (d (1 + log tft, d)
4. The score is 0 if none of the query terms are present in the document.

- **Calculation of idf [Inverse Document Frequency]**

  We define the idf weight of term t as follows:

$$\mathrm{idf}_t = \log_{10} \frac{N}{\mathrm{df}_t}$$

Important Points

1. N is the number of documents in a collection.
2. Alamban is a measure of the informality of the word.
3. Log "N / DFT] instead of [N / DFT] to" damp "the effect of ID.

- **Calculation of tf-idf weighting**

The tf-idf weight of a term is the product of its tf weight and its idf weight.

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \cdot \log \frac{N}{\mathrm{df}_t}$$

The result of this module generates a product amount for each document that will help evaluate the top ten news recommendation.

Module 4: Similarity Matching & Recommendation

This module will take the user's choice in terms of keywords and threshold values to decide the cutoff for the recommendation. The final group of news will be recommended as the final output.

5. CONCLUSION

This research work analyze and place the need of modern News Recommendation system which is based on user choice. This research also recognize that machine learning approach could help to classify the news into different fields and TF-IDF could help to find the similarities in news and also decides the related news. A model of 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.

6. REFERENCE

[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]. 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

[7]. 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.

[8]. N. Chen, S. C. Hoi, S. Li, and X. Xiao, "SimApp: A framework for detecting similar mobile applications by online kernel learning," in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 2015, pp. 305- 314.

[9]. 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.

[10]. 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.

[11]. Z. Zhou, Z. Cheng, L. Zhang, W. Gaaloul, and K. Ning, "Scientific workflow clustering and recommendation leveraging layer hierarchical analysis," IEEE Transactions on Services Computing, vol. 11, no. 1, pp. 169-183, 2018.

[12]. 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.

[13]. L. Qi, X. Zhang, W. Dou, and Q. Ni, "A distributed locality sensitive hashing-based approach for cloud service recommendation from multi-source data," IEEE Journal on Selected Areas in Communications, vol. 35, no. 11, pp. 2616-2624, 2017.

[14]. S. Deng, H. Wu, J. Taheri, A. Y. Zomaya, and Z. Wu, "Cost performance driven service mashup: A developer perspective," IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 8, pp. 2234-2247, 2016.

[15]. S. Deng, L. Huang, Y. Li, H. Zhou, Z. Wu, X. Cao, M. Y. Kataev, and L. Li, "Toward risk reduction for mobile service composition," IEEE transactions on cybernetics, vol. 46, no. 8, pp. 1807-1816, 2016.

[16]. 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.

[17]. F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: introduction and challenges," in Recommender systems handbook: Springer, 2015, pp. 1-34.