HYBRID NEWS RECOMMENDATION SYSTEM USING TF-IDF AND ASSOCIATIVE CALCULUS

Rachana Naik[1], Abhishek Raghuvanshi[2]

Assistant Professor
Mahakal Institute of Technology (IT) RGPV, Bhopal MP, India

Abstract: The growing use of internet makes it essential for daily life. It plays important role in busy schedule to make it easy and simple. The bigger challenge of today is awareness of current affairs. Data mining techniques is the results of an extended method of analysis and merchandise development. Data mining takes this organic process on the far side retrospective knowledge access and navigation to prospective and proactive data delivery. Recommender systems is one of the biggest outcome of data mining gives more relevant and useful outcome. NEWS Recommender systems have created important space in daily routing life. Newspapers are essential to urge information concerning recent activity and general awareness. Varied solutions are developing to convert paper News system to digital news and become an excessive amount of standard. This paper has investigated the importance of news recommendation solution and effort to improve the performance of news recommendation using modified TF-IDF algorithm. Proposed solution is implemented using Java technology and evaluated on basis of computation time for different category. A BBC dataset has been used as data source for same.

Keywords: TF-IDF, News, BBC Dataset, Associative Calculus

1. INTRODUCTION

Data Mining is the kind of data extraction technique used by various algorithms to explore more similar and relevant content. It broadly classified into two way; Predictive analysis and Descriptive Analysis. Both techniques can help to explore more relevant and accurate results with efficient manner. Here

Subsequently, Recommendations systems provide intellectual apply supported user preference. Recommendation systems provide separate and specialized set of data. In recent years, net personalization has received abundant attention to assist net users with the matter of data overload.

This work observes that integration of recommendation system and data mining approach can help to extract knowledge based on user preference and trend of today's clustered. This solution not only recommends the interested results but also involve the trend interest and popularity factor.

News recommendation system offers assortment of relevant news, articles, and suggestions supported user interest. They will offer news supported news quality and visits. News ranking, priority, area, impact etc could also be the core logic behind any news recommendation system. The salient expectation from news recommendation state that recommendation system should be able to recommend multiple results based on similarity factor. For example if user is looking to find news related to politics for particular party, it should recommend all respective news belong to that word along with co-relevant news based on current trend.

This works aims to extract all relevant news based on user demand words along with the unseen news that belong to same category but high factor. An you tube view can be considered as the source of inspiration for same but in news section.

Recommendation system can use variety of solutions such content mining or collaborative filtering. Classification of Association rule implementation can be another solution. This work has considered TF-IDF as the base algorithm for content mining and document mapping has been done though customized algorithm.

In this paper, a hybrid approach has been implemented using TF-IDF algorithm and customized document mapping concept. Next section comprises the basic study of previous work and associated problem. Afterwards, Problem formulation, proposed solution and implementation and result analysis has been provided. This paper ends with the concluding remark and relevant references.

2. RELATED WORK

Michal Kompan and Maria Bielikova proposed a Recommendation System to extract relevant news according to user preferences. They have used Slovak News portal and gave a solution to recommend news based on content mining. This solution divide the news information into two sections named as article and user to generate personalize recommendations. This work is based on article similarity algorithm and uses title, title words in the content, Category, Keywords, Names/Places etc. At first they preprocessed the News article and then the recommendation were made based on the ratio of recommended and visited articles, and recommended but not visited articles. However this solution provides a good recommendation system but we found that they have not used some quality matrices like popularity of

news and relevancy of news. A Block Representation of their proposed solution is shown in figure 1.
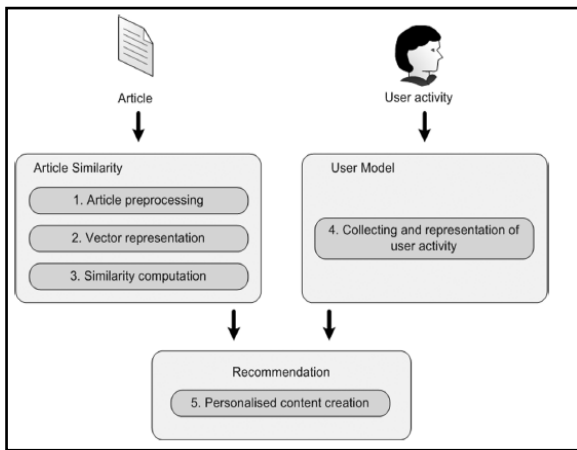

Figure 1: News Recommendation System

A brief description for the procedure of above explained solution is cited below;

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

Adnan et. al [2] explored an Fuzzy Logic based technique to overcome the drawbacks of TF-IDF algorithm. They observe that TF-IDF only explore the similarity at content level. It does not attempt to evaluate the sense of word along with the similarity with other words. They also explore that TF-IDF is based on the pot of words model so it does not consider semantics or co-occurrence position in text. It can be treats as the lexical feature based technique good to find equal position.

The study of conventional algorithm address that TF-IDF algorithm follows the concept which measure how many number of times that words appear in a given document. It also emphasizes to disconnect the prepositions and helping words before processing any task. Here, each word attempt to find relevance in a normalized data format which also add up to one. A formula to calculate the same is shown below;

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

RachnaMehrawat&Abhishek Sharma [] presented a survey of various existing solutions and derive problem formulation for news recommendation system. They also explore the importance of data mining approach and advantages of news recommendation system. They presented several relevant solutions and its importance for same. This work considered published problem as the guideline to derive the solution.

3. PROBLEM DOMAIN

The complete study observes that enhancement in data generation creating heavy load of data mining techniques. It becomes more complex when we need to extract very specific and perfect level of outcome. Previously, TF-IDF algorithms and content mining has been used for similarity extraction and relation finding.

Study observes that such algorithms only work on the concept of word similarity. Relationship with documents and involvement cannot be observed everybody has distinction perceptions and completely different reading feeling. It should vary as per user preference and job demand. Quality of content and impact of data is additionally necessary for user search.

A NEWS Portal is classified into different sections based on similarity of news articles and nature of news content. All sections have equal importance and completely different nature.

A Content-based News Recommendation system is projected by author [1] suffer with the issue of accuracy and relevance among documents. This works relies on article similarity rule and evaluated on non English database. So the work evaluation for English database with document mapping is expected.

The complete study observes that there is strong need to revised content mining approach from word based to document based. Subsequently, proposed solution should be able to evaluate accuracy and similarity for English word based database.

4. SOLUTION DOMAIN

News is the one of the important part of daily life. News may define as "Newly received or noteworthy information, especially about recent events". It can be state as "Information that is reported in a newspaper, magazine, television news program etc". It helps to make people updated and aware about the current affairs. Coverage of news may be subject of interest and people may like to read News from specific area or relevant topic.

The core parts of information mining technology are beneath development for many years, in analysis areas like statistics, computing, and machine learning. TF-IDF or collaborative filtering may be great option to implements the recommendation system but having certain scope of improvement. A similarity matching algorithm or association rule may used to implement the performance of the system.

Inadequate knowledge of search tool and large amount of data gives poor performance to retrieve or extract desire information. Recommendation systems offer intellectual practice based on user preference. Recommendation systems offer separate and specialized set of information. In recent years, Web personalization has received much attention to help Internet users with the problem of information overload.

The complete study conclude that web personalization in the field of News recommendation may help to improve the quality of content mining and recommend more useful and relevant piece of information.

The complete proposed solution is break down into three modules which are listed below;

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

Proposed algorithm of this work has been shown in figure 2.

```
Pseudo code 1: TF-IDF Algorithm
for(i=0 i<numnberOfUniqueWordsi++)
 for(j=0 j<numberOfDocuments j++)
tfidf = fij _ log(numberOfDocuments = ni)
for(s=0 s<numnberOfUniqueWords s++)
fijTemp                                            =
number_of_occurrences_ofword_S_in_the_document_J
tfidfTemp = fijTemp * log(numberOfDocuments / dfi)
summTfidf += (tfidfTemp)2
 end
 A[i,j] = tfidf/summTfidf
 end
end
```

Figure 2: Pseudo code 1: TF-IDF Algorithm

```
Pseudo code 2: TF-IDF Algorithm
D1={d1,d2,d3…..dn}
D2={d1,d2,d3…..dn}// Another Copy of D1
WDi=Wordlist of D1
WDn={wd1,wd2,wd3…wdn}
Ac={wd1||wd1,wd2,wd3…wdn}+{wd2||wd1,wd2,wd3…wd
n}…WDn
```

Figure 3: Pseudo code 2 for Similarity Matching

Proposed solution integrates both Pseudo code 1 and 2 into single module and proposed a hybrid approach for news recommendation.

This work uses BBC database of English language with the size of more than 2000 transaction has been used for testing purpose. In the first step and IR approach has been implemented using TF-IDF algorithm. Before implementing this lemmatization and Tokenization process, it has been implemented for smooth execution of proposed solution. After the successful implementation of TF-IDF, a document mapping algorithm has been implemented for similarity matching. Here, every word of one document is compared with another word of another document. For example,

considered two documents A and B. Suppose A has word I LOVE INDIA. And B document has text I LOVE MADHYA PRADESH. So this document will compare the every word of document A such I, LOVE and INDIA with every word of document B. So N8N comparison is expected in the proposed solution.

The complete solution would propose an integrated module of TF-IDF and similarity matching algorithm.

5. RESULT OBSERVATIONS

This work is implemented and evaluated using JAVA technology. Accuracy, Precision and Final Score parameters are used to evaluate the performance of proposed solution.

For evaluation purpose different categories has been observed from database, which are listed below;

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

To explore more accurate and relevant results, this work consider set of five words from each category and supply as the input source. The complete input data is shown in Table-1.

Table 1: Data Input

Category	Word
Business	Company
Politics	Government
Technology	Security
Entertainment	Award
Sports	Stadium

Table 1 has used as the input source and every iteration expect that final score of relevant category should be higher than others. To accomplish this desire a comparative result analysis approach has been performed where outcome has been compared with expected category.

The comparative study of all evaluated results is shown in Figure 3.

Table 2: Comparison of Previous Algorithm

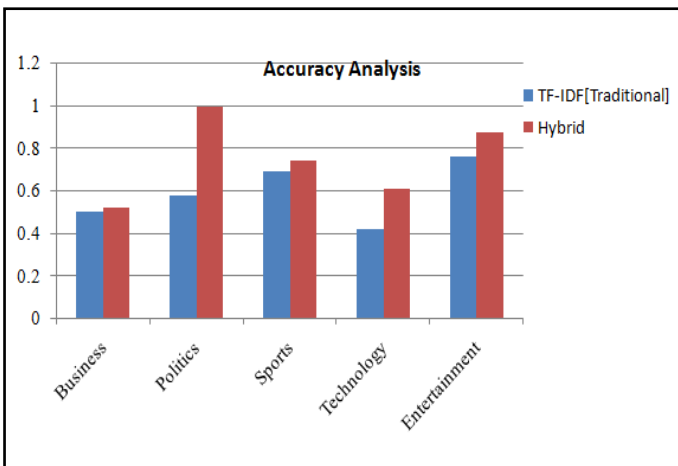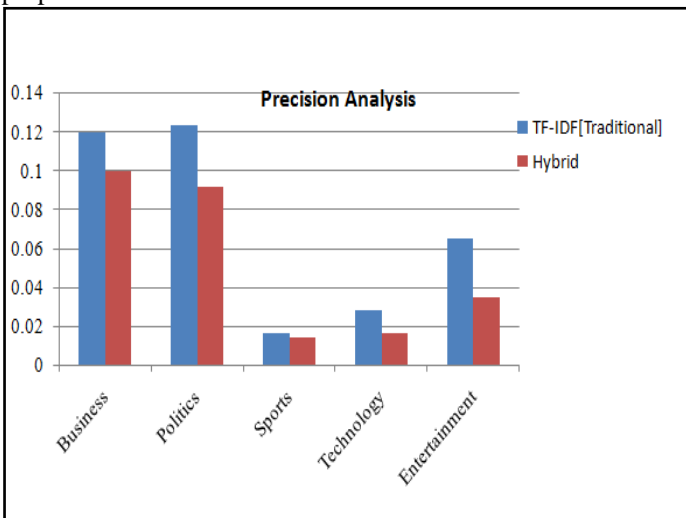| BBC-Dataset	TF-IDF	Previous Work	Proposed
Best Accuracy	0.8	0.587	1.0

Figure 3: Accuracy Comparison between traditional and proposed

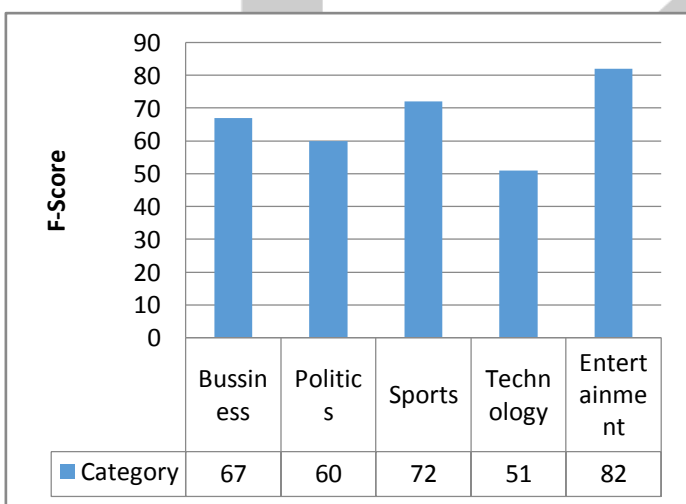Figure 4: Precision Comparison between traditional and proposed

Figure 4: Final Score of Hybrid algorithm

system won't solely observe the news content on user preference or quality basis however conjointly refine article on priority and impact basis. Planned system can facilitate to refine common and effective news content in step with user want.

The future scope of proposed solution is it can be tested and expanded with other news dataset such TOI or the hindu. This work only concentrate to improve accuracy further final score improvement is also expected.

REFERENCE

[1]. NeerajRaheja, V.K.Katiyar," International Journal of Computer Science Issues" Vol. 11, pp-2, 2014.
[2]. BahramAmini, Roliana Ibrahim, MohdShahizan Othman, "International Journal of Computer Science & Engineering Survey",vol 2,pp-3,2011
[3]. MinsukKahng, Sangkeun Lee, Sang-goo Lee, Ranking in Context-Aware Recommender Systems,pp-65-66, 2011.
[4]. Ch.Nagini, M.SrinivasaRao, Dr. R.V.Krishnaiah, International Journal of Engineering Research & Technology, Vol. 2,pp-701-704,2013.
[5]. Michal Kompan, M_ariaBielikov, Content-based News Recommendation,pp-1-12.
[6]. GediminasAdomavicius, Young, Kwon Improving Recommendation Diversity Using Ranking-Based Techniques, pp-1-33.

6. CONCLUSION

Recommender systems have created important progress in recent years and plenty of techniques are planned to boost the advice quality. However, in most cases, new techniques are designed to boost the accuracy of recommendations, whereas the advice diversity has usually been unnoticed. Planned