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

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

sujeetared8@gmail.com* , bmctmohitcs@gmail.com**
Mtech Scholar BM College of Technology Indore**

ABSTRACT

News Recommendation system has created a big space in daily routine life. News papers are essential part of daily life and we always look to collect important and sensitive information in single place. Varied solutions are developing to convert paper News system to digital news and become an excessive amount of standard. This paper will give the idea to generate important and sensitive news based on user choice from huge amount of news collection and it will also help to find news relation and keep track on related articles based on content relation. Revised TF-IDF algorithm with collaborative algorithm based mining framework has been developed and tested on BBC data based of accuracy, precision and recall. A Java tool has been developed for same.

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

1. INTRODUCTION

Now a days it is very difficult to find desired information over the internet. Sometimes users get irrelevant information and large amount of information gives poor performance to extract the desire information. So here we present a recommendation system which offers separate and specialized set of information. And it helps to prevent the users to get wrong information.

This conclude that Recommendation System is helps to prevent the users to get wrong information and it gives the good performance to extract the desire information. And it provides pre specified knowledge based on information. It is useful in different-different fields like news, marketing, shopping and many more.

News recommendation system is one of the crucial system for us. It has various news, articles and these all are based on the current situation.

News recommendation system offers collection of relevant news, article and suggestions. These all are totally based on user preferences.

Recommendation system uses different-different technologies. It can be classified in two categories:

1. Content based system.
2. Collaborative filtering.

Content based system:

This system explores the properties of the items recommended.
Ex:- If a user has watched many horror movies then he will get a recommendation of horror movies.

Collaborative filtering:

This system recommended items which are similar to user. This system explores all the requirements of users and the items they have.

2. Existing Work

There are two people named Michal Kompan and Maria Bielikova proposed a Recommendation System to extract the related news according to user preferences. They gave a solution for the recommended news. That solution is a Slovak news portal. It divides the news into two categories:

1. **Article.**
2. **Personalized Recommendation.**

This work is based on the article title, content, category and recommendation by users.

The block diagram of Recommendation system shown in fig 2.1.

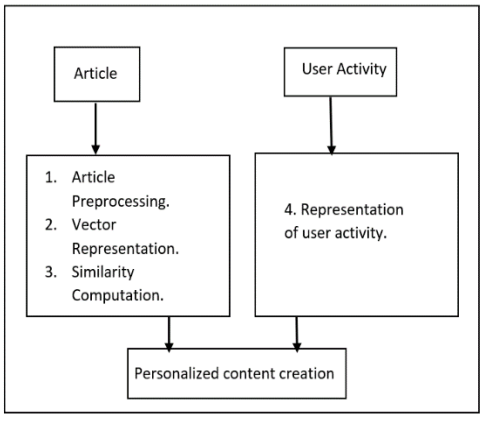

Figure: 2.1 News Recommendation System

In the above process they use some mathematical formulas for compute number of keywords, article relevance, name and places and many more from article content.

The complete study concludes that there is need to develop popularity and uses based recommendation tool to gather popular and important relevant news at one place.

This is extremely important when a recommendation list is created. Our method respects each of these types. We can easily redefine our similarity with simple changing the weights for vectors parts and adjust it for various recommender methods.

3. Problem Domain

There are lots of issues but one of the major problem is irrelevant information over the internet and overloading of information. Nowadays hectic schedules and poor knowledge of technologies are also major issues that affect the whole system.This problem becomes more sensitive and crucial when we try to extract current affairs and news from newspapers and online sources.

A News Paper divided into various sections like city, sports, editorial, international, national, entertainment etc. All thissections have equal importance and different user followers. Some time there may be possibility that, they may consist 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.

The complete study concludes that there is need to develop popularity and uses based recommendation tool to gather popular and important relevant news at one place.

4. Proposed Solution

Proposed solution is a recommendation system which offers separate and specialized set of information. And it helps to prevent the users to get wrong information. There is a problem of information overloading which overcome with the help of web.

The whole study conclude that News Recommendation system has issueslike the quality of content mining and recommend more useful

and relevant piece of information. These all are improve with the help of web personalization.

The complete scenarios are developed into four Modules which are listed below;

1. Dataset
2. TF-IDF Algorithm.
3. Similarity matching and computation approach.
4. News Recommendation.

Module 1: Dataset

There are some Sample Datasets are considered in the form of .csv format. A brief Description about dataset is shown in Table 3.1.

Table 3.1: Dataset Description

Dataset Name	Size	Number of rows	Description
BBC	6.00 mb	2234	It contains different-different news, articles. This belongs international level of news
Signal Media	6.02 mb	2267	It contains lots of images and videos related to news.
Sports	5.54 mb	2254	It contains various sports news which can be also international.
Politics	7.02 mb	3012 mb	It contain various political news which can be national or international.

Module 2: TF-IDF Algorithm

This module is used for the implementation of an algorithm for retrieval of information. To retrieve crucial information from the dataset using the data mining approach. The TF-IDF Algorithm is an information retrieval algorithm which is based on occurrence of keywords in whole dataset as well as whole document. It also shows the time which takes during the occurrence of keywords.

At the end of this module top ten documents are recommended according to their frequency matching value.

Module 3: Similarity matching and computation approach

As we seen previous module woks on top ten documents which are recommended according to their frequency matching value. And this module also works on those top ten documents and try to find the similar documents and similar content of dataset using computation approach. Suppose dataset has 100 documents then user match all the documents with top ten document using computational approach.

It also matches every word of document with every words of top ten documents. So, if a document has 200 words and another document has 150 words so total word matching will be 3, 00,00.

This module helps to estimate the similarity computation of top ten documents with another existing documents to explore more relevant news suggestion and recommendation.

Module 4: News Recommendation

This module integrates all the documents. Firstly, this proposed system creates a dataset that is in an excel file. Now it explores all the files and does lemmatization and removes stop words. It also follows similarity matching computation. It computes all the calculation using TF-IDF such as low weight, DF, length normalization, and product sum. After the calculation it explore the top ten

news recommendation then integrate the news and lastly, it generate top 15 news recommendation. All process are shown below in figure 4.1:

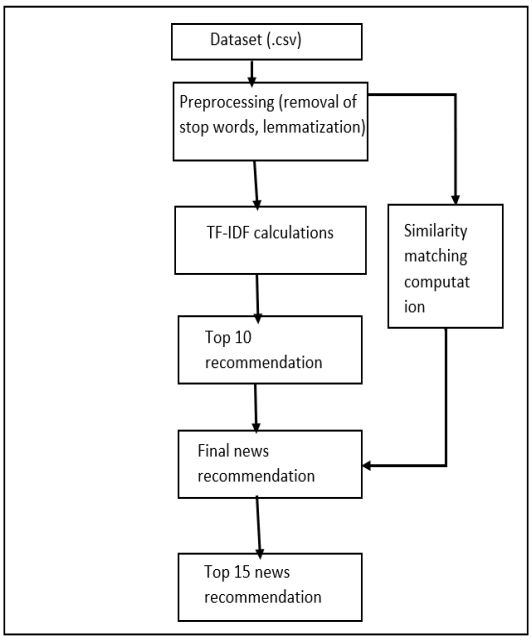

Figure 4.1: News Recommendation

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

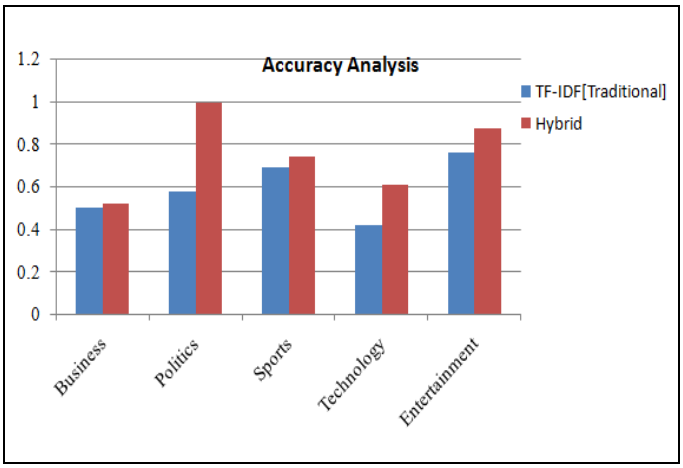

Figure 3: Accuracy Comparison between traditional and proposed

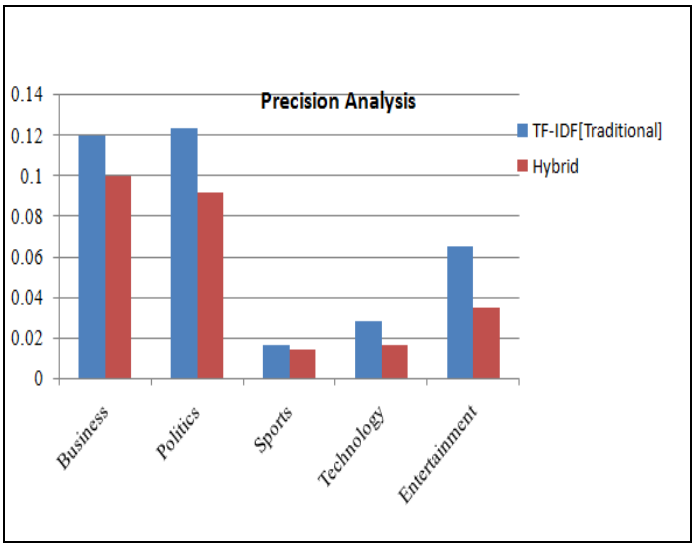

Figure 4: Precision Comparison between traditional and proposed

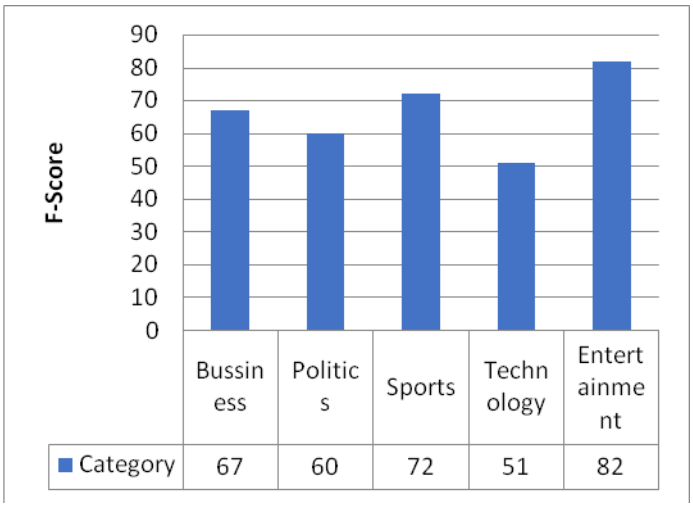

Figure 4: Final Score of Hybrid algorithm

5. Conclusion

The complete work concludes that the proposed solution gives better performance for all recognize words.There are lots of issues but one of the major problems is irrelevant information over the internet and overloading of information. This proposed solution overcome all of these problems. The proposed solution has four modules and each module has its own importance.The first module is for dataset creation and the dataset will be in the form of a CSV file. The second module is to retrieve information from the dataset. The third module is for matching documents. The fourth module is the integration of all three modules.

REFERENCE

[1]. Neeraj Raheja, V.K.Katiyar," International Journal of Computer Science Issues" Vol. 11, pp-2, 2014.

[2]. Bahram Amini, Roliana Ibrahim, Mohd Shahizan Othman, "International Journal of Computer Science & Engineering Survey",vol 2,pp-3,2011

[3]. Minsuk Kahng, Sangkeun Lee, Sang-goo Lee, Ranking in Context-Aware Recommender Systems,pp-65-66, 2011.

[4]. Ch.Nagini, M.Srinivasa Rao, Dr. R.V.Krishnaiah, International Journal of Engineering Research & Technology, Vol. 2,pp-701-704,2013.

[5]. Michal Kompan, M_aria Bielikov, Content-based News Recommendation,pp-1-12.

[6]. Gediminas Adomavicius, Young, Kwon Improving Recommendation Diversity Using Ranking-Based Techniques, pp-1-33.