International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

# Personalized Health News Recommendation Based on User Interests

**Hitesh Bhatia[1], Yash Gurnani[2], Devesh Hinduja[3]**

**Project Guide: Mr. Amit Singh[4]**

[1-4]*Dept. of Information Technology, Vivekanand Education Society's Institute of Technology, Maharashtra, India*

---

**Abstract –** *Online news reading has become very popular these days because the web provides access to various news articles from thousands of sources all around the world with variety of regional, local, national and international news in all the languages. A key challenge of these news applications is to help users find the articles that are interesting to read which differ from user to user. In this paper, we present our research on developing personalized news recommendation system based on user's interests regarding health related topics. Based on the predefined interests that we ask from users, we developed a Bayesian system for predicting users' current news interests from the activities of that particular user and the news trends. We combine the information filtering mechanism using learned user profiles to generate personalized health news recommendations.*

***Keywords*: Keyword Extraction, Health News Recommendation, News Scraping, Information Filtering, YAKE! Algorithm, News Personalisation**

## 1. INTRODUCTION

News reading has changed with the advance of the World Wide Web, from the traditional model of news reading via physical newspaper subscription to access to thousands of articles/sources via the internet. News aggregation websites, like Google News and Yahoo! News, collect news from various sources and provide an collected view of news from around the world. A critical and major problem with news service websites is that the never-ending volumes of articles can be overwhelming to the users. The challenge is to help users find news articles that are interesting to read based on their particular interest. This technique is known as Information Filtering. Based on profiles of each and every user with their interests and preferences, system recommends news articles that maybe of interest or value to the user. So, our main task is to aggregate news articles according to user interests and creating a "personal newspaper" for each user.

Nowadays, recommendation of personalized content is becoming the most popular area for many researchers. The main aim of recommendation is to provide meaningful suggestions to users for particular items based on the user's interest and the profile. News is the important part in day to day life. There is a tremendous increase in volume of digital news, articles and the choice for people to read news as per his/her interests has also increased significantly. Thus there is a need for a system that will accurately give suggestions to the user depending on its interest.

An accurate profile of users' current interests is very important for the success of information filtering systems. Some systems do it manually for every user to create and update profiles. Sometimes users don't like this and it takes a burden sometimes and only very few are willing to do it. Instead, some systems construct profiles automatically from users' interaction and activity with the system.

In this paper, we describe our research on developing a personalized heath news recommendation for every user based on user profile learned from user's activity on the system. For our system, we are scraping news from various websites and also gathering news using newsapi in the domain of health. After getting news from our scrapers, we are extracting summary from the article. And from that summary, we are generating keywords from every article using YAKE! algorithm. After doing so, we are checking users' activity based on the news clicked and the keyword extracted from that certain news article. Based on their activities, we are recommending news to each and every user with the help of keyword extraction associated with clustering of interested keywords which acts as interests.

This paper is organized as follows. Section 2 explains the purpose of the proposed plan. Then the Literature Survey in Section 3. Then the three main functions of the proposed program are described which tells us of text preparation, corpus clustering and news recommendation in Section 4, 5 and 6 respectively. The flow diagram is made in Section 7 where the working is explained. Finally, a conclusion is made in section 8.

## 2. OBJECTIVES

Following are the main objectives:

- Developing the recommendation model.
- Improving the accuracy of the model.
- Providing the health news to the readers based on the language and interests.
- Recommending the health news to the readers as per their profile, interests, user activity, etc.

## 3. LITERATURE SURVEY

Multiple papers were studied and their findings are summarized in this section. This section includes papers studied before and during the development of the project.

The papers helped in gaining insight into existing solutions, possible ways to optimize algorithms and facilitate the selection of algorithms based on their performance.

In Neural news recommendation with negative feedback (Oct 2020, Springer) the authors did classification on news as positive or negative feedback according to the user based on session time and dwell time.

In A Framework for Benchmarking Stream-based News

Recommenders (Jan 2020, IEEE Xplore) the authors classified news based on the user clicks, items and sessions which were recommended on four categories - 1.Clicks per item 2.Clicks per user 3.Session per user 4.Clicks per session.

In Neural News Recommendation with Long- and Short-term User Representation(Feb 2019, IEEE) the authors first divided and classified the news according to the session time and reading time and then based on the above classification they created two categories Short Term User - which doesn't spend much time on certain article and Long Term User – which spends more than usual time on certain news article.

In Contextual Hybrid Session-based News Recommendation with Recurrent Neural Networks(Oct 2019, IEEE Access) the authors created a hybrid of LSTUR(Long and Short Term User Representation) and user content with 3 global factors which were – 1.Popular Topics 2.Seasonality of news 3.Breaking/Trending Events.

## 4. DATASET DESCRIPTION

For getting the dataset, we have scraped the health related news from various websites naming –

1. Healthline Media
2. News18 India
3. BBC News
4. Hindustan Times
5. ScienceDaily LLC
6. WebMD

After scraping the articles from the above websites, we are considering total of 6 attributes which our scraper will collect and they are as follows -

Headline

Summary

imageurl

datetime

article

articleurl

After getting our dataset, we tried to extract keywords from our article's summary for each and every article. We performed exploratory data analysis on the data. Firstly, we normalize the data by removing the stopwords and preprocessing it to get a cleaned and normalized text corpus.

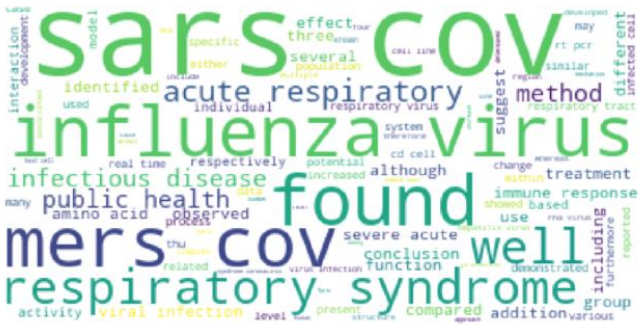


Fig 1.Word Cloud of our Corpus

## 5. TEXT PREPARATION

Text in the corpus needs to be transformed to a format that can be interpreted by the machine learning algorithms. There are 2 parts of this conversion — Tokenization and Vectorisation. Tokenization is the technique of converting the continuous text into a list of words. Then the list of words is converted to a matrix of integers by the process called as vectorisation. Vectorisation is also called feature extraction. For text preparation we use the bag of corpus which will be ignoring the sequence of the words and only will consider word frequency.

We used 4 different kinds of algorithm for our model preparation and try to extract the most accurate keywords from news articles for better accuracy.

### A. TF-IDF and Scikit-Learn

TF-IDF stands for Text Frequency Inverse Document Frequency. The importance of each word increases proportionally to the number of times a word appears in the document which will be the word's frequency (Text Frequency - TF) but is offset by the frequency of the word which is inverse (Inverse Document Frequency - IDF). Using the TF-IDF weighting scheme, the keywords are the words with the highest frequency hence will have highest TF-IDF score.

So, for a summary –

*"Migraine patients consider red wine the principal alcoholic trigger, but studies show that other types of alcohol are just as likely the culprit."*

Following are the list of keywords extracted using TF-IDF -

```
===Keywords===
principal
red wine
wine
trigger
alcoholic
consider
migraine
red
alcohol
type
```

Fig 2.Keywords extracted using TF-IDF

This algorithm doesn't help us to find accurate keywords from the above summary.

### B. Gensim Implementation of TextRank Summarization Algorithm

Gensim is a Python library module which is free and is designed to automatically extract semantic words from documents. The gensim is implemented on the popular TextRank algorithm. It is an open-source topic assessing toolkit, implemented in the Python programming language, using NumPy, SciPy and optionally Cython for performance.

So, for a summary –

*"On World Mental Health Week, Aamir Khan shared how mental and physical hygiene hold same importance and also gave small tips on how to beat stress."*

Following are the list of keywords extracted using gensim-

```
===Keywords===
mental
hygiene hold importance gave small tip beat
week aamir
```

Fig 3.Keywords extracted using gensim

Gensim doesn't help either to find accurate keywords from the above summary.

### C. RAKE-NLTK

RAKE short for Rapid Automatic Keyword Extraction algorithm, is a not depended on any domain for keyword extraction. It works by analyzing the frequency of word appearance and its occurrence with other words in the text. RAKE usually doesn't originally print keywords in order of score. But it returns the score and the extracted keywords.

So, for a summary –

*"On World Mental Health Week, Aamir Khan shared how mental and physical hygiene hold same importance and also gave small tips on how to beat stress."*

Following are the list of keywords extracted using rake-nltk-

```
===Keywords===
list include deepika padukone
various celebrities coming
several academy awards
reportedly lost 20kgs
best visual effects
best film editing
best adapted screenplay
best picture
best director
best actor
tom hank
seen playing
ranveer singh
professional front
next seen
mental health
manisha koirala
kareena kapoor
karan johar
insecure side
```

Fig 4.Keywords extracted using rake-nltk

We used RAKE for the whole context of news from certain article, but the results are very bad. So, we move onto our next keyword extractor algorithm which will be used as our model with better accuracy.

### D. Yet Another Keyword Extractor(YAKE)

It is an unsupervised approach for Automatic Keyword Extraction using Text Features.

YAKE! is a unsupervised automatic keyword extraction method which is also light-weight and it rests on text statistical features extracted from single documents to select the most important keywords of a text. This system doesn't require training on any certain set of documents, neither it depends on dictionaries, external-corpus, size of the text, language or domain.

So, for a summary –

*"Migraine patients consider red wine the principal alcoholic trigger, but studies show that other types of alcohol are just as likely the culprit."*

Following are the list of keywords extracted using yake –

```
===Keywords===
('migraine', 0.15831692877998726)
('culprit', 0.15831692877998726)
('patient', 0.29736558256021506)
('red', 0.29736558256021506)
('wine', 0.29736558256021506)
('principal', 0.29736558256021506)
('alcoholic', 0.29736558256021506)
('trigger', 0.29736558256021506)
('study', 0.29736558256021506)
('type', 0.29736558256021506)
```

Fig 5.Keywords extracted using yake

YAKE gives better results than all of our above used algorithms for keyword extraction.

So, we are going to use yake for our model preparation in our system for better and accurate keyword extraction from news articles.

## 6. CORPUS CLUSTERING

Using the algorithm Mini Batch as a Classifier Stochastic gradient descent is the dominant method used to train deep learning models.

Gradient descent is an optimization algorithm which is mostly used for finding the weights or coefficients of machine learning algorithms.

It works in such a way where the model makes predictions on training data and using the error on the predictions to update the model in such a way as to reduce the error.

The goal of the algorithm is to find model parameters (e.g. coefficients or weights) which will reduce and henceforth minimize the error of the model on the training dataset. It does this by making certain changes to the model that moves it along a gradient down toward a minimum error value. Hence, it got its name as "gradient descent."

Batch gradient descent is a slight variation of the gradient descent algorithm that calculates the error for each example in the training dataset, but only updates the model after all training examples have been evaluated.

One completion of cycle through the entire training dataset is called a training epoch.

Mini-batch gradient descent is another variation of the gradient descent algorithm that divides and splits the training dataset into small and equal batches that are used to calculate model error and update model coefficients.

During implementation it may choose to sum the gradient over the mini-batch which will further reduce the variance of the gradient.

Mini-batch gradient descent tries to find a balance between the robustness of stochastic gradient descent and the efficiency of batch gradient descent. It is the most common implementation and very helpful way of using of gradient descent and it is also used in the field of deep learning.

## RESULTS

Since, from the above four extraction algorithms, Yake is the most accurate algorithm giving better keywords and successful results.

Using the Principal component analysis (PCA) to decompose the data in project it to a lower dimensional space, we get clusters arranged with the help of corpus clustering with relative frequency keywords extracted from our summary.

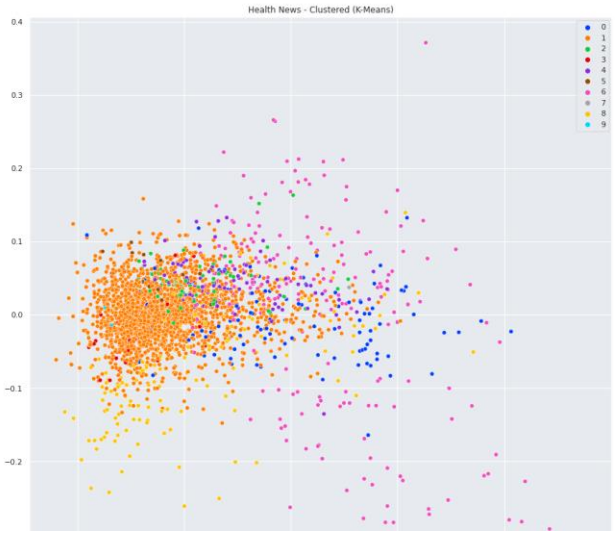


Fig 6.Cluster of our Keyword Extracted Corpus

Figure 6 shows us the K Means Cluster of our Yake algorithm generated keywords.

And figure 7 shows us the visualization of K means Cluster of Yake algorithm generated algorithms in 3D for a better visualization.

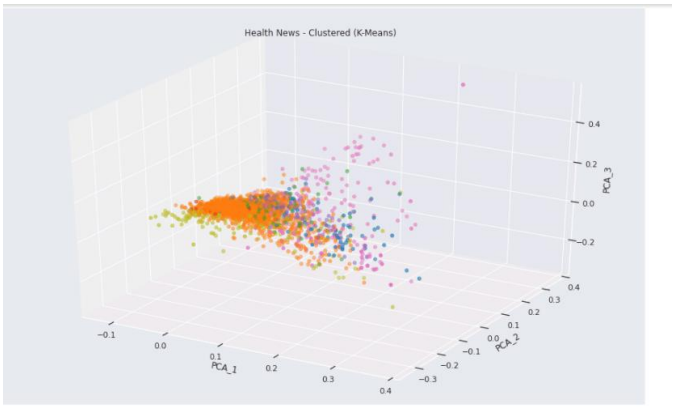


Fig 7.Cluster in 3-Dimension

Hence, we generate clusters of each summary in the dataframe and figure 8 shows the size of each cluster.

| cluster | |
|---|---|
| 0 | 103 |
| 1 | 2208 |
| 2 | 75 |
| 3 | 26 |
| 4 | 99 |
| 5 | 25 |
| 6 | 215 |
| 7 | 1 |
| 8 | 159 |
| 9 | 14 |

dtype: int64

Fig 8.Size of each Cluster

## 6. RECOMMENDATION MODEL

After extracting keywords from every article and from that extracted keywords list we created a corpus and did some clustering. With this, we got clusters which are frequent and relevant with their similar keywords.

Now, in our system, every user will have a predefined interests selected for themselves. With the help of that interests, we have created a parent domain and each domain contains a sub-domain(For e.g. a parent domain named Fitness will have sub-domain as gym, yoga, exercise, etc as their involved keywords).

So, every user will have his predefined interests and based on our keywords extracted we map each user with their following interests. Then, every user will get news articles based on their selected interests with the help of keywords extracted.

Now, coming to recommendation of news –

The system has set a threshold of 3, this tells us that; if some user is reading certain news so the keywords used in that article will add up to user interest. If some keyword is read thrice, which means a user is reading various news and a count is been set for keywords coming up in that article for every user.

And, if a new keyword comes up thrice, which is a threshold of 3, we will add that keyword in user's interests list.

Since, the content is too much, and new news gets added every time, we check the user activity based on the time spent on certain article and keyword gets automatically updated if it crosses the threshold.

With this technique, the system keeps a check on user activity and recommends the news articles which acts as their new interests and user profile gets updated automatically.

## 7. Flowchart for the Proposed System

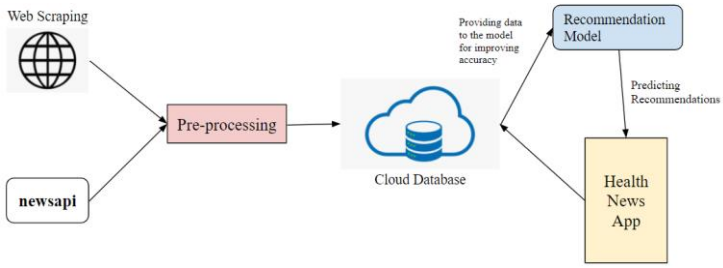


Fig 8.Proposed flow diagram of our system

The overall flowchart of the system is shown in figure above. We extract news from our scrapers and newsapi and we preprocess and extract keywords from every article and create a cluster of every similar and frequent keyword. Then we add our news content to our database with keyword extracted from our model. Then after user sign up, we ask user their predefined interests and start showing them news based on interested articles. As per user activity in the system, our recommendation model works and dynamically updates in our database for every user. Based on this updation, it starts reflecting on our application based on the activity and profile of every user.

## 8. Conclusion

In this paper, we present our research on developing and effective information filtering mechanism for health news recommendations in a large-scale dataset which is mentioned above. We first showed user the trending health related news based on the language chosen and location. Also, we ask every user for their initial pre-defined interests and show news based on that selected interests. Based on these findings, we check user activity and select them into two parts – the genuine interest news from every user and trending/local news for all users. A Bayesian framework is proposed to model a user's genuine interests using her past read history and predict her current interests by combining her genuine interest and the new keywords added in user

profile. The method for predicting user's interests was used in news information filtering, and it was combined with the existing collaborative filtering method to generate personalized news recommendations. We conducted this experiment on our android application and the system was working fine. We got the good results and also the recommendation of health news articles was working just fine.

**ACKNOWLEDGEMENT**

We would like to thank the above mentioned websites to let us scrape the news articles from their respective pages. And we would also like to thank newsapi.org which helped us getting news articles with the help of free trial API. We would also like to acknowledge our Project guide Mr. Amit Singh for providing necessary guidance.

**REFERENCES**

[1] https://newsapi.org/

[2] https://www.healthline.com/

[3] https://www.news18.com/

[4] https://www.hindustantimes.com/

[5] https://www.bbc.com/news/health

[6] https://www.sciencedaily.com/

[7] https://www.webmd.com/

[8] Personalized News Recommendation Based on Click Behavior - Jiahui Liu, Peter Dolan, Elin Rønby Pedersen

[9] Contextual Hybrid Session-based News Recommendation with Recurrent Neural Networks(Oct 2019, IEEE Access)

[10] Neural News Recommendation with Long- and Short-term User Representations(Feb 2019, IEEE)

[11] A Framework for Benchmarking Stream-based News Recommenders(Jan 2020, IEEE Xplore)

[12] Neural News Recommendation with Negative Feedbacks(Oct 2020, Springer)