Personalized News Recommendation System

[2]Senevirathna W.M.M.P, [3]Perera M.S.S, [1]Dr.Pradeep Abeygunawardhana[4]Karunarathne R.N.N.B, [5] Senavirathna B.M.A.M,

[1]Department of Information Technology Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
[1]pradeep.a@sliit.lk, [2]marlonprabhath@gmail.com, [3]shohanperera03@gmail.com, [4]nilupul.n@gmail.com, [5]

ashaniseneviratna@gmail.com

Abstract---With the development of network information technology, a lot of news comes into the view of Internet users. We have entered the "information overload" era. So how to find useful information becomes more and more important. People feel it very difficult to read all the news articles and find the information what they want. It takes lot of time as well as man effort. Personalized news recommendation is a kind of technology to find the news that users want to get urgently. It is based on the browsing history of many users. In this system we have identified four main components such as extraction and classification, aggregation, summarization and recommendation.
Collaborative filtering recommends an item to a user, based on the references user's preferences for the target item or the target user's preferences for the reference item. **Content-based filtering** ***uses data about a user's search results, browsing history and/or purchase history to determine which content to serve to the user. Through using these new algorithms, we can generate recommendation lists and realize the personalized news recommendation.***

Keywords—
Personalized recommendation, Collaborative filtering, Content based filtering, Hybrid recommendation algorithm, TextRank algorithm, Word2vec, WordNet, Multinomial Naïve Bayes, Support Vector Machine, Random Forest

I. INTRODUCTION

The main purpose of this Research paper is to illustrate recommending the classified, summarized and most accurate news to the users and this will be based on the user preferences such as their interest and behaviors. Personalized news Recommendation is a more important thing in nowadays because with this high technology people can't waste their time, money and their effort too. Also the information overloading is one of the serious problems nowadays since information is generating in a rapid rate with the advent of internet. When it comes to the e-news context this problem remains the same. There is enormous volume of news articles from numerous portals on the web. These contain gigantic amount of news articles from all around the world added daily at a rate of hundreds, even thousands or more per hour. This prohibits a difficulty for the access to the right information and users must spend a lot of time manually shifting out useful or relevant information. So recommending correct news at the correct time according to their specifications without delaying may cause fewer difficulties to the users.

II RELATED WORK

A. NEWS EXTRACTION AND CLASSIFICATION

There are increasing number of e-news websites, therefore lots of methodologies have been proposed for extracting and classifying the e-news article text contents of e-news websites.

Some methodologies use the semi-automatic or manual example based wrapper learning to extract the e-news article text content from e-news websites. Shinnou et al. proposed an extraction wrapper learning method which expected to learn the extraction procedures and that could be applied to e-news pages from other numerous e-news portals [27]. Fukumoto et al. concentrated on subject shift and presented an approach for extracting key paragraphs from text document, which discuss the similar event [28]. That process uses the results of event tracking which begins from sample text documents and discovers all following documents. Though, if an e-news site uses many diverse layouts in the e-news pages, the learning process consumes excessive time and the precision will be low. Reis et al. presented a calculation of the edit distance between two trees for the full automatic web news article content extraction [29]. Webstemmer is a web crawler which analyses HTML layout and automatically extract key text of an e-news pages excluding advertisements, banners and navigation links. It examines each page layout in particular e-news website and discovers the location of the main text. All the analysis could be done with very little human interaction in an automatic way.

There are several e-news classification approaches. Several studies related to text classification exist in the overview. Jiang et al. proposed a text classification approach based on a modified K-nearest- neighbor algorithm. It is combined with a constrained one pass clustering algorithm [30]. Uguz et al. The proposed algorithm for reducing the number of features using information gain feature selection approach. For feature selection the genetic algorithm and principle component analysis are applied. Then for the classification k-nearest neighbor algorithm and decision tree algorithm will be used. This approach is efficient, but could be improved by the introduction of few more powerful classifiers or an ensemble of classifiers [31]. Lee et al. suggested an improved support vector machine classification approach that uses a Euclidean distance function which is reported to have low impact on soft margin parameter C and the implementation of kernel function [32].

B. NEWS AGGREGATION

There was a research "An Unsupervised Content Based News Personalization using Geolocation Information" [25] done by Khumukcham Robindro, Kshetrimayum Nilakanta, Deepen Naorem and Ningthoujam Gourakishwar Singh. The main purpose of this research was to get rid of redundancy of the news articles which describe the same content in different sources. Clustering approach has been used to remove redundant news and represent a representative cluster using well known k-means clustering algorithm. The news article with the highest amount of content has been selected as the representative cluster. Clustering has been performed in a way that only the relevant clusters would be selected to increase the recommendation quality of news articles. Further, recommendation of news has been enhanced by extracting details related to locations of the users in order to recommend local news more accurately. Natural Language Processing can be stated as a possible way to improve the identification of exact locations of the news.

In this approach the term frequency – inverse document frequency (tf-idf), is used to convert a textual document into Vector Space Model. The tf-idf method is a very interesting way to evaluate how important is a word in a document.The first step is applying preprocessing techniques such as removing stop words, stemming. After preprocessing a dictionary should create by using all the terms present in documents. All words in the dictionary are taking as the dimensions of the vector space. The number of dimensions of the vector space is equal to the number of words in the dictionary. The tf-idf vector is calculated for each document separately by calculating tf-idf values for each dimension for a particular document.

C. NEWS SUMMARIZATION

This section discusses some of the existing automatic text summarization approaches. Hans Christian, Mikhael Pramodana Agus, Derwin Suhartono et al. [22] discussed a Term Frequency-Inverse Document Frequency (TF-IDF) based text summarization approach. TF-IDF was used as a measure to score the sentences which is based on the word frequencies which reflects how important a word is to a document in the corpus. They had only used nouns and verbs of the text considering that they bring the important details about the text. Then a score was assigned to each sentence by taking the sum of the TF-IDF values of every noun and verb in the sentence. Then the sentences were arranged in the descending order of their scores and the final summary was generated using only the top ranked sentences. The amount of sentences extracted for the summary depends on the compression rate. The major limitation of this approach is that the accuracy of the results was low since they have used only one measure, the TF-IDF to extract the important sentences from the text.

Sumya Akter, Aysa Siddika Asa, Md. Palash Uddin, Md. Delowar Hossain, Shikhor Kumer Roy, and Masud Ibn Afjal-etal. [23] designed a multi document extractive summarization system using the k-means clustering algorithm. In this approach the word scores were given based on the TF-IDF measure and then the sentence scores were assigned by summing the term frequencies of words in the sentence with its position. If any cue word is present in the sentence the sentence score was incremented by one. Then the sentences were ranked in the descending order of their scores and two clusters were initialized taking the maximum sentence score and the minimum sentence score as initial centroids. Then the Euclidean distance from each sentence to the two centroids were calculated and the sentences were assigned to the cluster which has the minimum distance. Then new centroid values for each cluster were reassigned and the same process was repeated until the centroid values won't change. Finally the top sentences from each cluster were taken to form the final summary. The limitation in this approach was that it consumes lot of time and effort since defining the k value at the beginning is tricky and the most

optimal k value can be gained only by the trial and error process

D. NEWS RECOMMENDATION

Personalized recommendation has become an essential service of major sites. News portal, to provide various types of news, is a traditional internet service, but there is still a big gap on personalized news service compared to today's thriving e-commerce site. Users of the consulting class websites is very wide, if better able to tap the potential of user's interest and achieve news recommendation accordingly, it can produce greater social and economic value. More importantly, personalized services can enhance the adhesion of users by its good user experience.

Personalized recommendations core technology is the study and application of recommendation algorithms and collaborative filtering algorithm is the most mature research, but also the most widely used recommendation algorithm. Such as the famous news site Digg, it tried to recommender systems in his home. The site calculates the interest similar between users according to the user's search history, and then recommend similar users' favorite articles to target user. According to the website statistics, users' search behavior significantly become more active after using the recommended system. Recommendation systems can be categorized into two sections. Those are personalized and non-personalized (popularity-based recommender system). Non-personalized recommendation means without considering individual user's preference, system provide general recommendation to the users. When we login to the news portal it provides most popular news items around the world. These popular items based on age, geography, sex, count of purchases, feedback etc. [4], [7]. Based on these parameters, system calculates the mean of the news rating of all the users and list down the news articles according to their mean value. This also called as a "stereotyped recommender system". But there is a problem with this system, when there are less number of ratings available mean value will be less accurate. So, this kind of a system provide less confidence.

III METHODOLOGY

A. NEWS EXTRACTION AND CLASSIFICATION

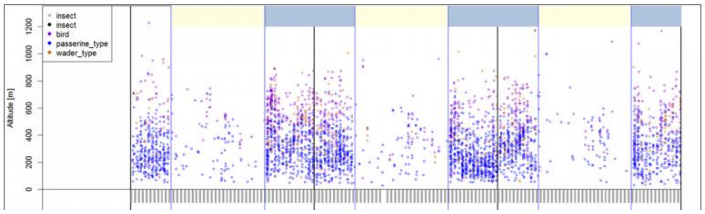


Figure. 1. Extraction and Classification Architecture

Figure 1 shows the overall module design. There are three main phases in this system. There are e-news extraction module and e-news classification module.

A. e-news Extraction Module

E-news extraction module uses state-of-the-art tools, which could length with functionality to meet the defined requirements. This approach combines two advanced methodologies. At the beginning of the process user provides root URLs or RSS feeds into the system. Web crawling and extraction are the next stages of the process. Newspaper, python library is used for scraping and Feed Parser is used to read RSS-feeds. These days most of the e-news websites provide RSS feeds. If there is RSS-feed available, it will be used for scraping e-news articles. Newspapers' automatic e-news article scraper is used for other sites. Scrape from the RSS feed first, since the data was much more reliable when gather through the RSS feed. To feed system from RSS feeds and root URLs JSON file will be used. Therefore, it will be convenient to add or remove e-news websites. FeedParser will be used to load the RSS feed of the e-new website. Build the structure for the data by constructing dictionary Newspaper. An article dictionary is created to store records for each e-news article. Then *Newspaper* library is used to scrape the content of the links from the RSS-feed. Finally, the data object should save. Published date of the article will be checked in extraction. It will give more accurate results.

B. Pre-processing Module

The next phase is pre-processing. The pre-processing procedure is significant in text processing as well as text classification, because it supports to sum up an article effectively by removing redundant words and conforming each word to its root form, helping the classifier to recognize the article more easily and efficiently. In pre-processing we remove noise form data, which are irrelevant to this process such as advertisements, user comments, etc. And do Transform cases, Text Tokenization, Stop word removal, Word stemming and Lemmatization.

1. Transform Cases: Transform case use to convert all the news items into lowercase letters. It helps get an effective outcome from other pre-processing stages.

2.Text Tokenization: The tokenize technique breaks raw e-news strings into sentences, then breaks those news sentences into words and punctuation, and after applies a part of speech tag. This approach eliminates white spaces, tab, new line, etc. The token is then normalized. Use NLTK tool kit for Text Tokenization.

3. Stop Word Removal: E-News becomes a rapidly growing communication medium where different type of users is involved in. As a result, irrelevant textual data, abbreviations, irregular expressions and infrequent words can be created. To reduce that noise of textual data is essential for accuracy of the sentiment analysis. Hence, to produce quality data set we are removing such stop words from the web comments by using pre-compiled stop word lists or stop word identification in NLTK package. Stop word removal is done by default the set of English language stop words from NLTK is used.

4. Word Stemming: Stemming is applied to find out the root or stem of a word. There are sets of rules to apply. This is one of the most important steps in the pre-processing process. Stemming reduce the time consuming and space required and that increase efficiency of the classifier. In this approach S-Stemmer, Lovins, Porter, and Husk Stemmer will be used.

5. Lemmatization: Lemmatization is the procedure of looking up a single word form from the range of morphologic affixes that could be applied to specify tense. First need to recognize the WordNet tag form based on the Penn Treebank tag, which is returned from NLTK's standard pos_tag function. If the tag with 'N', 'V', 'J' or 'R' then can properly identify if it's a noun, verb, adjective or adverb. We then use the new tag to look up the lemma in the lexicon. The WordNetLemmatizer looks up data from the WordNet lexicon and do Lemmatization on the news items.

The classification module consist of feature extraction and ensemble classification

1. Feature Extraction

TF-IDF model is used to extract features. As the dataset, we used BBS news dataset which contains 2225 news articles with class labels. The TF-IDF contains Term Frequency and Inverse Document Frequency. Term frequency summarizes how often a given word appears in a document. Inverse document Frequency considers no of documents as well. Therefore, it downscales words that appear a lot across documents. We extracted 14788 features using TF-IDF.

2. Ensemble Classification

$$log\frac{\#docs}{1+\#docs\ using\ word}$$

Figure. 2. Ensemble Classifier

Ensemble classifier will be used for this classification. Assemble together several different algorithms or several different models to create an ensemble learner. It gives low error and low over fitting than standalone classifiers. The theory of combining classifiers is suggested as a novel direction for the improvement of the performance of individual classifiers. Support Vector Machine (SVM), Neural Networks, Naive Bayes classifier, Decision trees, Discriminant Analysis, Nearest Neighbours and Random Forest algorithms are the most powerful classification algorithms. Therefore, by evaluating these techniques, constructed ensemble classifier using SVM, Random Forest algorithms and Multinomial Naïve Bayes.

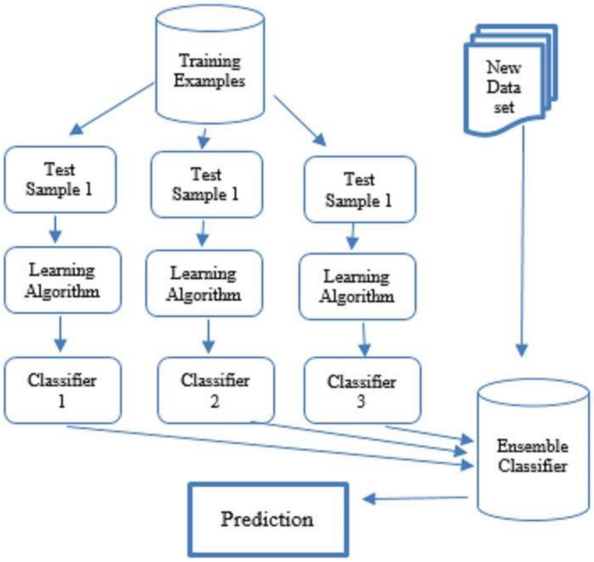


Figure. 3. Ensemble classifier

SVM: This linear classification method can be used for multiclass classification other than the binary classification. SVM which is doing the classification using linear decision boundaries is called as linear SMV and as well as with the little enhancement of the algorithm SVM can be modified for nonlinear classification which uses the non-linear decision boundaries. SVM is a supervised learning algorithm and for a given set of training data this algorithm generates an optimal hyperplane which can use to categorize new data items. SVM is commonly recognized to be a more accurate algorithm.

Random Forest Algorithm: Random forest algorithm is an inbuilt ensemble classifier, it consists of two main stages. There are random forest formation and make predictions from the produced random forest. Initially arbitrarily select “k” amount of features from total “m” amount of features. (k<<m) Then define the node “d” using the best split point among the “K” features. Subsequently divide the node into order nodes using best split. Next that repeat above steps until “I” number of nodes has been created. Then create the forest by repeating all above steps “n” number of times.

Multinomial Naïve Bayes: This classifier is suitable to classify discrete features. It is a probabilistic classifier based on text features. Naïve Bayes classifier can be trained very efficiently by requiring a relatively trivial quantity of trained data.

Neural networks: Different forms of neural network approaches have been applied to document classification. This research used multilayer perceptron which is more sophisticated than single layer perceptron. This network contains an input layer, two hidden layers and single output layer. Tensorflow library is used to build neural network and a Bag of Word model is used to produce a unique list of words. This model is used as a tool for feature generation. Then built neural network and train the model using the training data set.

Ensemble classifier: Used individual trained models for classification. First extract features from news article using TF-IDF, then feed those features into saved model vectors. Next get probabilities from each model. Those probabilities are used to generate weighted average probability for each class. Than aggregate probabilities from different classes and get the maximum value class as predicted class for particular e-news item. Below shows probabilities, for example e-news article.

Below e-news item gives following probabilistic results
RF- Random Forest Algorithm
MNB- Multinomial Naïve Bayes Algorithm
SVM-Support Vector Machine Algorithm

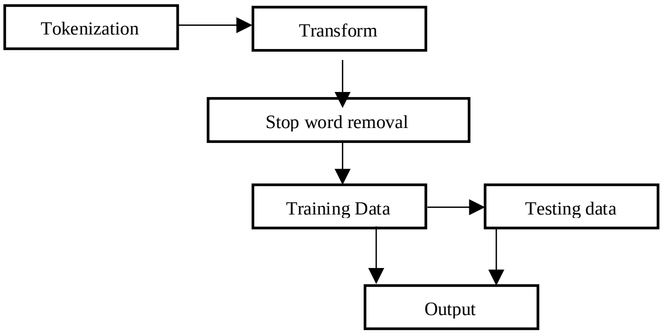


Figure. 4 an E-news item

Ensemble probability = P_RFN+oP o_fM clNasBsi+fPier_sSV M

P_RF=Probability value from Random Forest Algorithm
P_MNB= Probability value from Random Forest Algorithm
P_SVM= Probability value from Support Vector Machine

Ensemble those individual classifiers using probabilities.

B. NEWS AGGREGATION

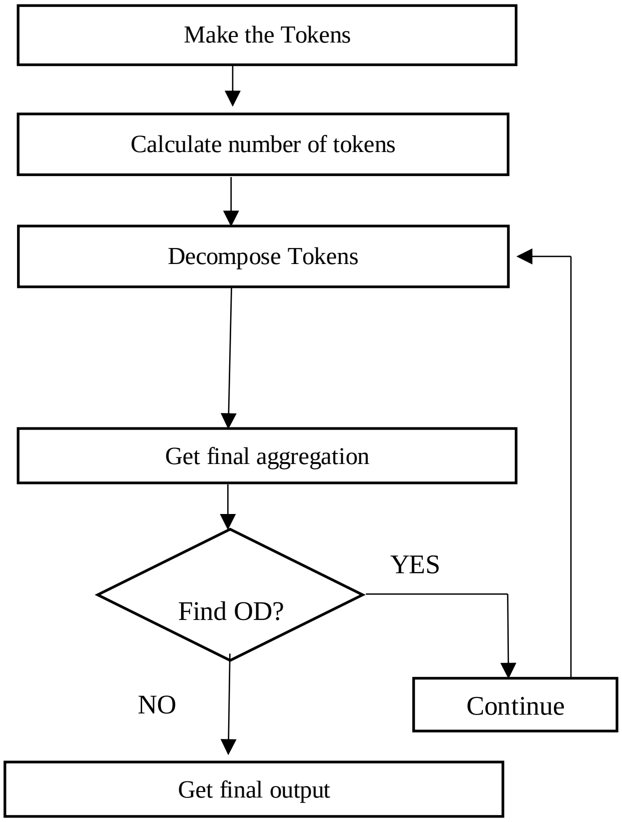


Figure. 5 an E-news item

Firstly pre-process to extracted contents and do stemming. Stemming is the process of finding base form of a word. Stemming is important because same word can exist in different forms in different documents [16]. *NLTK.stem* function has used for stemming. After pre- processing create a dictionary from those stemmed data. Create a corpus using bag of words technique to get the number of times each word occurs. Next count is created for each article using TF-IDF method for clustering the news articles text documents should transform into numerical values [17]. These numerical are called as the features.

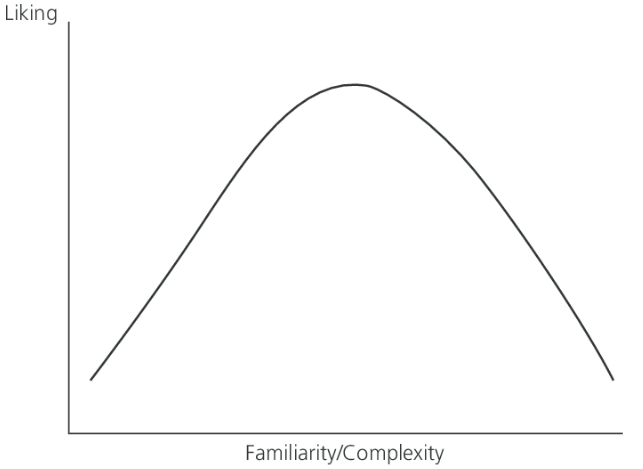


Figure 6 -Curve analysis

By using similarity measurement get similarity score by compare each and every articles with one articles. That technique repeat for every articles with one article. To do that we generate comparison and ignoring same article comparing and the article previously compared with that article by using bellow algorithm. In here *"i"* is the checking document, "j" is the comparing document, and "contents" is compare article saved list. "extracted_data"is all article saved list that get from database, "add_doc" is the function that generate similarity with each articles use similarity technique in.

```
contents = [ ] for i in
range(len(extracted_data)-1):
for j in range(i+1,len(extracted_data)):
        contents (extracted_data[j]['content'])
add_doc(extracted_data[i]['content'],contents)
    del contents[:]
```

NEWS SUMMARIZATION

From this module get e-news cluster with store news articles same title and generate an extractive summary. First the news articles in the cluster admit to pre-processing one by one. Then individual summary will generate for each article .then all the summaries collect and generate intermediate summary by the system relevant to given e news cluster.

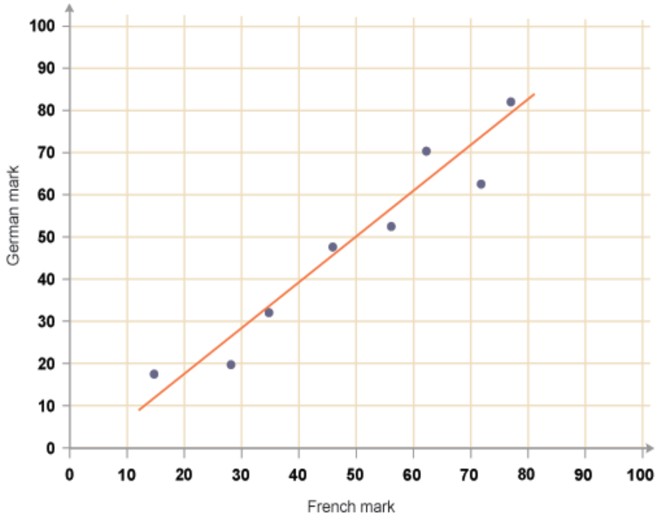


Figure 7 : Linear graph analysis

The system build corpus by using document 2 bag of word (doc2bow) module then summarized that corpus by building the hashable corpus to keep the summarization quality warning and minimum requirement will be check. After that graph will be build using PageRank algorithm [19]. So the sentences were represented as nodes and the similarities between the sentences were represented as edges in the sentence similarity graph. Then each node in the graph was assigned a score [20] using the PageRank algorithm which is an indication of how significant a particular sentence in the summary. Then get graph edge weight using Best Matching 25 Algorithm (BM25) [21].

In here Given Q have q_i keywords. *f*(*D*,*qi*) is q_i's term frequency

in document D, IDF(q_i) is Inverse Document

Frequency weight of the query term q_i

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)},$$

to compute IDF value use this formula

$$\mathrm{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

Where N is the total number of document in collection, *n(q$_i$)* is the documents containing count.

I V. RESULTS AND DISCUSSION

Compare with that weight filter the important content from the sentences. The summary for one article is generated .then collect all the summaries and do page ranking and apply BM25 methods to build up final summary. The summarization process involves the major stages as described below.

A. Elimination of Redundancy

Since salient sentences were extracted from multiple documents in the cluster there were duplicates of information which conveys the same idea. These redundancies needed to be removed because otherwise it will lead to poor system generated summaries. The redundancies of sentences were identified by considering sentence similarities in three major perspectives namely the lexical similarity, syntactic similarity and semantic similarity between the sentences. Separate algorithms were used to identify those similarities and by utilizing these similarity measures between the sentences, these redundancies were removed from the aggregated summary and then the final summary was generated.

1) *Removal of Lexical Redundancy*: Lexical similarity between sentences were identified as overlap of words between the sentences. In order to identify the overlap of words between sentences the jaccard similarity measure was used. Then it is judged that the sentences are lexically dissimilar by the condition that the jaccard similarity between the two sentences is less than the threshold of 0.6. The jaccard similarity measures the similarity between two sentences as the amount of word overlap normalized by the union of the sets of words present in the two sentences [26].

2) *Removal of Syntactic Redundancy*: In order to identify the syntactical relationships between sentences, 2-gram models were generated to each sentence in the sentence pair and then computed the 2-gram probability. If the 2-gram probability value is greater than zero it was concluded that there is kind of a syntactical similarity between the sentences in the sentence pair.

3) *Removal of Semantic Redundancy*: Although there were some syntactic similarity between sentences sometimes there were no semantic similarity between those sentences. Therefore it was not possible to simply remove syntactically similar sentences without checking it for the semantic similarity. Therefore the semantic similarity identifies sentences which have the same meaning. The semantic relationships between sentences were identified as an average score from the two approaches namely word2vec based semantic similarity and WordNet based semantic similarity.

A. Content-Based Filtering Algorithm

Content based algorithms are used for content-based recommendation engines. It's based on user's past behaviors and interests [7], [8]. Google and Wikipedia are examples for these kinds of systems. The basic idea of this is keeping keywords of the past articles and provide recommendations based on that. But biggest problem in here is large pie of information provide some difficulties to provide recommendations. As an example, if someone search for "The University Culture", there will be large number of documents containing the keywords "The" and same as for the "Culture" as well [9], but collaborative filtering algorithm is most widely used algorithm for the news recommendation. News portal like Digg uses this kind of technique as their recommendation engine [5].

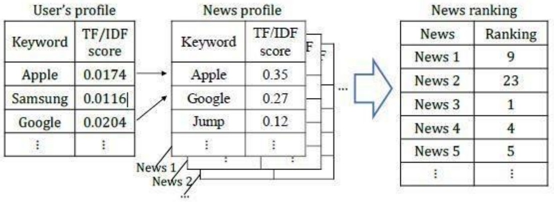
News content is often represented using vector space model. A well-known method is TF-IDF value of the corresponding keyword. Based on a user's history behaviors, the user profile can be created. For a newly published news articles, we can compute the similarity between user profile and the news articles by similarity functions (Jaccard Similarity or Cosine
Similarity).

$$cos(p_u,q_i) = \frac{}{\|p\ \|2\|q_i\|2} = \frac{\sum q_{ut}p_{it}}{\sqrt{\sum_t q_{ut}^2}\ \sqrt{\sum_t}} \quad {}_u p_u .\ q_i \quad {}_t \quad q_{i2t}$$

News Ranking

Created user profile can be used for the news recommendation. For each user, every news article set is evaluated to find the similarities between them. As a first step need to be crawled an upcoming news and then filtered noisy words and sentences to extract relevant information from the article. Second, calculate the TF-IDF scores and then stored in "News Profile". Last, these similarity scores which are in user's profile and news profile computed using cosine similarity function.

Following figure shows the brief understanding about above scenario. According to the rank of the news recommended to the user.



Keyword	TF/IDF score
Apple	0.0174
Samsung	0.0116
Google	0.0204
⋮	⋮

Keyword	TF/IDF score
Apple	0.35
Google	0.27
Jump	0.12
⋮	⋮

News	Ranking
News 1	9
News 2	23
News 3	1
News 4	4
News 5	5
⋮	⋮

Fig. 6. News ranking according to the TF-IDF score

B. *Collaborative Filtering Algorithm*

Collaborative recommender systems are based on collaborative filtering algorithm. It's based on nearest neighbor concept, it consists two major filtering concepts [1], [4]. Those are User-based filtering and Item-based filtering. User-based filtering means consider the similarities between users and Item-based filtering means it consider the similarities between items (news articles). To calculate how similar two users, use Karl Pearson's correlation formula [1], [10].

Improved Pearson Correlation Coefficient Formula News have the characteristic of strong timeliness; top news will get a lot of users to click and comment in a specific period. When the recommendation system analyzes the user's interests and calculated the similarity of users [11], two users reach a consensus over controversial news items more valuable than over the hot news. Visibly, the hot factor will seriously affect the recommendation system on mining interest of users, thereby affecting the personalized service provided to users.

Therefore, optimization of Pearson Correlation Formula by introducing the parameter of hot can reduce the importance of the popular news to finding similar users [12], improve the recommendation accuracy rate and enhance the user experience. The hot *hj* of news *j* is calculated to analyzing as follows;

$$h_j = \frac{\sum_{j=1}^{N} r_{i,j}}{N} \quad (h_j > 0)$$

Here, *N* represents the total number of users (including users who did not score on the news), *hj* represents the ratings of the user *i* to the news *j*. in calculating sum of rating, if the user *i* has no ratings record to news *j*, skip the user thus it can be seen that the more people score on the news *j* and the higher the score, the more popular the news. Each user's ratings on the news they visited be a vector, which is expressed as follows [5];

(r1,1, r1,2 ,… ,r1,n)

The ratings news set of user *x* signs as *jx* and the ratings news set of user *y* signs as *jy*. Union news set commented by users *x* and *y* signs as *jxy*, using the traditional Pearson

Correlation Coefficient Formula to calculate similarity between the user *x* and *y* as follows [5], [12];

$$\mathrm{sim}(x,y) = \frac{\sum_{j \in J_{xy}} (r_{x,j} - \overline{r_x})(r_{y,j} - \overline{r_y})}{\sqrt{\sum_{j \in J_{xy}} (r_{x,j} - \overline{r_x})^2} \sqrt{\sum_{j \in J_{xy}} (r_{y,j} - \overline{r_y})^2}}$$

Fig. 2 gives an example of user similarity matrix.

The improved Pearson Correlation Coefficient Formula used to calculate the similarity is;

$$sim^*(x,y) = \frac{\sum_{j \in J_{xy}} \frac{1}{h_j} (r_{x,j} - \overline{r_x})(r_{y,j} - \overline{r_y})}{\sqrt{\sum_{j \in J_{xy}} (r_{x,j} - \overline{r_x})^2} \sqrt{\sum_{j \in J_{xy}} (r_{y,j} - \overline{r_y})^2}}$$

As can be seen from the formula improved, the more popular the news, for the calculation of the similarity between user *x* and *y* smaller role.

C. *Hybrid News Recommendation Module*

Most recommendation systems consist only one or few recommendation algorithms. That's not reliable and not as much as accuracy for recommendation engines. So, we propose a Hybrid Recommendation System, consist with Content-based filtering, Collaborative filtering and Location aware personalization with user preferences. User profiles are used for tracking the user's long-term interest and short-term interest. Individual user profiles can be divided into two parts, those are static and dynamic user profiles. Static user profiles use for storing users' long-term interest and dynamic user profiles are used store users' short-term interest. Static profiles are constructed during user signup and dynamic profiles are created when user using the system [8].

Due to vast content of information provided by the online mass media, it's necessary to have a powerful database such as MySql. Mysql is a Sql database work as a fast and real-time data provider. It's important to have accurate and real-time news update for user to read information. Click frequency of the article *i* and category *j* relationship can show as below.

clicks*i* = total number of clicks made by one user on article *i*;

$$f_{click}(i, j) = \frac{\sum_{users} clicks_i}{\sum_{users}\sum_{i=1}^{n} clicks_i}$$

Users input query is matched with the snipped stored inside of the article database. When user search some news and try to get articles, it helps to provide best Content-based recommendation for the user. Related users are identified using User-based collaborative filtering and similarities between articles are calculated using Item-based filtering algorithm.

V. CONCLUSION AND FUTURE WORK

In conclusion this provide the design and implementation of an e-news recommendation system with automatically summarized news which are collected from several existing e-new sites by classifying and aggregating. Those news items are recommended to users as their favors and behaviors. The system is evaluated based on accuracy and user satisfaction. At present this system exist only for selected area, in future this will be available for many several areas whole over the world. plan to build a mobile application to make easy for users and recommend the news matching for user location.

.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Dr. Pradeep Abeygunawardhana for her continuous support, motivation, and valuable suggestions. Her guidance helped me in writing of this research paper and implementation of the project.

.

REFERENCES

[1] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. T. Riedl, “Application of Dimensionality Reduction in Recommender System -- A Case Study,” presented at the IN ACM WEBKDD WORKSHOP, 2000.

[2] S. Jiang and W. Hong, “A vertical news recommendation system: CCNS #x2014;An example from Chinese campus news reading system,” in *2014 9th International Conference on Computer Science Education*, 2014, pp. 1105–1114.

[3] Y. Wang and W. Shang, “Personalized news recommendation based on consumers' click behavior,” in *2015 12 th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2015, pp. 634–638.

[4] N. Jonnalagedda and S. Gauch, “Personalized News Recommendation Using Twitter,” in *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2013, vol. 3, pp. 21–25.

[5] S. Liu, Y. Dong, and J. Chai, “Research of personalized news recommendation system based on hybrid collaborative filtering algorithm,” in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, 2016, pp. 865–869.

[6] J. Sun *et al.*, “A Novel Approach for Personalized Article Recommendation in Online Scientific Communities,” in *2013 46 th Hawaii International Conference on System Sciences*, 2013, pp. 1543–1552.

[7] B. R. Cami, H. Hassanpour, and H. Mashayekhi, “A content-based movie recommender system based on temporal user preferences,” in *2017 3 rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS)*, 2017, pp. 121–125.

[8] K. Veningston and R. Shanmugalakshmi, “Personalized location aware recommendation system,” in *2015 International Conference on Advanced Computing and Communication Systems*, 2015, pp. 1–6.

[9] A. Tomar, “Content Based Recommender System in Python,” *Ankur Tomar*, 14-Aug-2017.

[10] N. C. Hoan and V. T. Nguyen, “Advance Missing Data Processing for Collaborative Filtering,” in *Computational Collective Intelligence. Technologies and Applications*, 2012, pp. 355–364.

[11] G. S. Sadhasivam, K. G. Saranya, and E. M. Praveen, “Personalisation of News Recommendation Using Genetic Algorithm,” in *2014 3rd International Conference on Eco-friendly Computing and Communication Systems*, 2014, pp.
23– 28.

[12] B. Jueajan, K. Naleg, L. Pipanmekaporn, and S. Kamolsantiroj, “Development of location-aware place recommendation system on Android smart phones,” in *2016 Fifth ICT International Student Project Conference (ICT-ISPC)*, 2016, pp. 125–128.

[13] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, “Collaborative Filtering Recommender Systems,” in *The Adaptive Web*, Springer, Berlin, Heidelberg, 2007, pp. 291–324.

[14] D. Werner and C. Cruz, “A Method to Manage the Precision Difference between Items and Profiles: In a Context of Content-Based Recommender System and Vector Space Model,” in *2013 International Conference on Signal-Image Technology Internet-Based Systems*, 2013, pp. 337–344.

[15] P. Suppasert, R. Pungprasert, K. Putkhaw, and S. Tuarob, “Newsaday: A personalized thai news recommendation system,” in *2017 6th ICT International Student Project Conference (ICT-ISPC)*, 2017, pp. 1–4.

[16] M. Anjali and G. Jivani, “A Comparative Study of Stemming Algorithms.”

[17] R. Kumar *et al.*, “Web Document Clustering and Ranking using Tf-Idf based Apriori Approach.”

[18] C. C. Aggarwal and C. Zhai, “Chapter 4 A SURVEY OF TEXT CLUSTERING ALGORITHMS.”

[19] K. Kumar and F. Donewell Mukoko, “PageRank algorithm and its variations: A Survey report,” *IOSR J. Comput. Eng.*, no. 1, pp. 2278–661.

[20] T. Sri, R. Raju, and B. Allarpu, “Text Summarization using Sentence Scoring Method,” *Int. Res. J. Eng. Technol.*, pp. 2395–56, 2017.

[21] “Practical BM25 - Part 2: The BM25 Algorithm and its Variables.”

[22] Christian, H., Agus, M. P., & Suhartono, D. (2016). *Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF).* ComTech: Computer, Mathematics and Engineering Applications, 7(4), 285. doi:10.21512/comtech.v7i4.3746

[23] Akter, S., Asa, A. S., Uddin, M. P., Hossain, M. D., Roy, S. K., & Afjal, M. I. (2017). *An extractive text summarization technique for Bengali document(s) using K-means clustering algorithm*. 2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR). doi:10.1109/icivpr.2017.7890883

[24] Na, L., Peng, X., Ying, L., Xiao-jun, T., Hai-wen, W., & Ming-xia, L. (n.d.). *A topic Approach to Sentence Ordering for Multidocument Summarization*

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation”, *Journal of Machine Learning Research 3*, pp. 9931022, 2003.

[26] Pal, A. R., & Saha, D. (2014). *An approach to automatic text summarization using WordNet.* 2014 IEEE International Advance Computing Conference (IACC). doi:10.1109/iadcc.2014.6779492

[27] H. Shinnou and M. Sasaki, "*Automatic extraction of target parts from a Web page*.," IPSJ SIG Notes, Vols. 2004-NL-162, pp. 33-40, 2004.

[28] F. Fukumoto and Y. Suzuki. *Detecting shifts in news stories for paragraph extraction. In TheProceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, 2002.

[29] D. de Castro Reis, P. B. Golgher, A. S. da Silva, and A. H. F. Laender. *Automatic Web news extraction using tree edit distance. In The Proceedings of the 13th International Conference on World Wide Web*, pages 502–511, 2004.

[30] Jiang, S., Pang, G., Wu, M., Kuang, L.: *An improved k-nearest-neighbor algorithm for text categorization*. Expert Systems with Applications 39(1), 1503–1509 (2012)

[31] Uğuz, H.: *A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm*. Knowledge-Based Systems 24(7), 1024–1032 (2011)

[32] Lee, L.H., Wan, C.H., Rajkumar, R., Isa, D.: *An enhanced support vector machine classification framework by using euclidean distance function for text document categorization*. Applied Intelligence 37(1), 80–99 (2012)