Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 5 – No.1, May 2016 – www.caeaccess.org

News Classification using Neural Networks

Gurmeet Kaur
Chitkara University
India

Karan Bajaj
Chitkara University
India

ABSTRACT
Classification of data in data mining is used to determine the category to which data belongs. Classification is required in order to search and manage data in databases. In this era of information, news are accessible very easily as news are available through online sources. This becomes a necessity to classify such data as news articles pose a great influence on various sections of our lives. This paper presents a system for the classification of news articles based on artificial neural networks and have compared the results with the previously used techniques for classification.

General Terms
News Classification, Algorithms, Feature Selection

Keywords
Classification, Stop words, Word Net, Text Mining, and Stemming.

1. INTRODUCTION
With the increasing amount of information or data stored in electronic format there is a need of powerful means for analysis and interpretation of such data which could be useful in decision-making process. The last few years have seen a growth of research interest in the development of textual data management techniques. These techniques are required as the textual data is increasing with the passage of time and such techniques help in performing indexing and retrieving such growing text data. Data mining is a powerful technology with great potential to predict future trends and behaviour. Data mining performs three basic operations and these are exploring the data, finding patterns in data and performing predictions. Text mining is used for the finding of information that was previously unknown by extracting such information from unstructured data. Now-a-days this un-structured data is growing rapidly. The un-structured data mean, the type of data in which the text is occurring in a natural free form or a sequence that may include word and sentence ambiguity. A few pre-processing methods are required in order to extract useful patterns and information from the unstructured text. This paper presents the classification of news fetched from BBC network.

News documents sometimes contains never-before-seen information. So news requires dynamic classification and discovery. As classification must be done using sparse training data, it pose problem for standard classification techniques. This paper presents algorithms for classification of news categories that are shown to be highly effective.

A large number of news related to various categories such as sports, technology, entertainment, music, and politics can be seen over the internet these days. Each user can see such types of news on internet. If the user is interested in news related to a particular section then one has to go at that option and then one can see that news by clicking at that news. This is a very time consuming process. If it is possible to display news for user as per his choice then it will be a better option. As the number of website providing news on the internet has increased, it has become very difficult for the user to get the news of his own interest. As a result it is mandatory to filter the news sentences in various categories so that users can access them easily. To resolve this a classification system has been proposed.

This paper will continue as follows. First, in section 2 a literature survey and related work will be examined. Then, in section 3 the process of category classification will be given. Next, in section 4 experimental results are given. Finally, in section 5 concluding remarks are made and future work discussed.

2. RELATED WORK
Dr. R. R. Deshmukh, Mr D. K. Kirange[1] has presented a news personalisation system that aimed at advocates news to the users according to their interests as per predefined in the user's profiles. In their paper the authors have presented a SVM based system for classification of headlines and for posting news to users as per their choices.

Huajiao Li1, Wei Fang, Haizhong An, Xuan Huang[2] presented a mechanism to combine statistics, word segmentation, complex networks and visualization to analyse headlines' keywords and words relationships in online Chinese news and had tracked the news evolution fastly.

Jinyan Li, Simon Fong, Yan Zhuang,Richard Khoury[3] have evaluated many classification algorithms and a few filtering schemes and combining all have performed hierarchical classification.

Mazhar Iqbal Rana, Shehzad Khalid, Muhammad Usman Akbar[4] have discussed text classification process, various classifiers, and a lot of feature extraction methodologies but all in context of short texts i.e. news classification based on their headlines.

Hai-Tao Zheng, Bo-Yeong Kang, Hong-Gee Kim[5] have utilized semantic relationships, ontologies such as WordNet used to improve clustering results. WordNet-based clustering methods mostly on single-term analysis of text; they do not perform any phrase-based analysis.

C.Ramasubramanian, R.Ramya[6] have defined the drawbacks of the existing stemmers and have used certain constraints to improve their efficiency.

Zach Chase, Nicolas Genain, Orren Karniol-Tambour[7] have analysed the performance of a binary classifier approach and have built a learning model that works similar to the way human will classify articles topics.

Dasa Munkova, Michal Munka, Martin Vozar[8] had worked in determination of necessity of carrying out of data pre-processing steps in e-documents.

3. CLASSIFICATION PROCESS

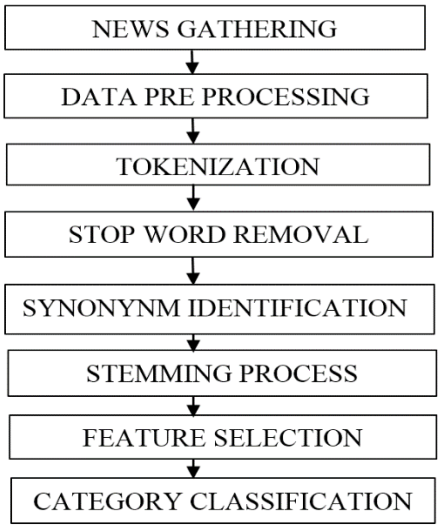

Fig1: Classification Process

The news classification process is achieved through various steps. These steps involves data gathering, Data pre-processing, Feature Selection, Implementation of proposed algorithm and comparing results with the previously existing algorithms.

4. NEWS GATHERING

It is not possible for human beings to keep account of all the current events and they rely primarily on the media view and interpret current events. As a result media is having a considerable influence on human lives. In the past the media reported about the current events

through newspapers, radio and television by delivering the news at regular intervals and times throughout the day. However in the past few years' news delivering websites have emerged as a medium through which individuals can obtain information on current events. Comparing online news with the previous news delivery media, these are delivered and updated at a much faster rate and people can have greater control on the way and time on which they can receive news. In our research the news have been gathered news from the online news channel BBC news.

5. DATA PRE-PROCESSING

The data that is obtained from the BBC website. Each obtained news consists of following documents: news headline, news description, link, and metadata like author and publishing date. The data that is obtained may be an incomplete, noisy and inconsistent. Such data makes the pre-processing of data a necessity as no results could be obtained from an incomplete information. Data Cleaning provides us unified date formats, noise free data, correct inconsistent data After applying the pre-processing news along with its headlines were extracted. Extraction is the method used to tokenize the file content into individual word.

6. TOKENIZATION

Tokenisation refers to breaking a sequence of text into symbols, phrases, words and all other meaningful elements into tokens. The aim of tokenisation is to explore the words in a sequence. At beginning textual data is only a block of characters. But while dealing with information retrieval system it is required to have a data set of words. So parsing is required to process tokenisation of documents. The main use of tokenisation is to identify the meaningful keywords.

7. STOP WORDS REMOVAL PROCESS

Generally some very common words that appears to be of very little value in helping in fulfilling user need are removed from the vocabulary completely. Such words are called stop words. Stop words are removed in order to save both time and space. Stop words are an integral part of information retrieval process. The removal of stop words increases performance and search results. The stop words needs to be removed for a reason since they provide no distinctive information for classification purpose. The stop words are generally language and task dependent. There are a few words that appear in very few documents are also filtered as they are not expected to be represent any category.

8. IDENTIFYING SYNONYNM

Using the word synonym can improve the classification process. For the synonyms detection and usage WordNet database is one of the best choice to be made. WordNet acts as a lexical database for English. It groups together English words into sets of synonyms known as synsets, it records a number of relations among these synonym sets and its members. It includes the lexical categories nouns, adjectives, adverbs, verbs but it pay no attention to prepositions, determiners, and other function words. It narrates how to find mutual informants between terms by using backdrop knowledge through WordNet.

9. STEMMING PROCESS

Stemming is a step of pre-processing Text Mining. It is very important step in most of the Information Retrieval systems. The purpose of stemming is to minimize different grammatical forms / word forms of a word like its noun, verb, adverb, adjective, etc. to its root form. The goal of stemming is to reduce inflectional forms and derivationally interchangeable forms of a word to a common base form. Text clustering, categorization requires stemming process as part of the pre-processing before actually applying any related algorithm.

Most of the times it has been noticed the linguistic variants of words have similar semantic interpretations and can be weighted as equivalent for the goal of Information outrival applications. As the meaning is same but the word form is different it is mandatory to identify each word form to its base form. To achieve this a large number of stemming algorithms exists. Each algorithm attempts to convert the linguistic variants of a word like introducing, introduces, introduction etc. to be mapped to the word 'introduce'. Few algorithms may map them to just 'introduc', but that is not a problem considering all of them map to the same word form. So the key terms involved in a query or a document are mirrored by stems instead of by the original words. The idea is to reduce the processing time of the final output.

There exists many stemmers like Lovins stemmers, Porter Stemmers, Paice/Husk stemmers, Dawson stemmers, N-Gram

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 5 – No.1, May 2016 – www.caeaccess.org

Stemmers, HMM stemmers, YASS stemmers, Krovetz stemmers, Xerox stemmers. All the stemmers have their own pros and cons. In our research Snowball has been used which is a string processing language which creates stemming algorithms to be used for stemming purposes.

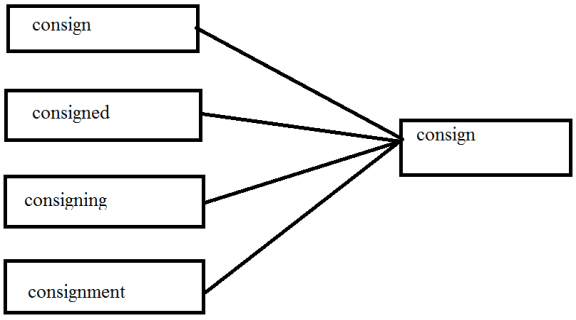

Fig 2: Stemming process

10. FEATURE SELECTION

When a huge amount of features are given and each of the features is a well-known descriptive word for each class, a lot of time may be consumed in classification and it may be possible that expected classification accuracy may not be achieved, and data may get over-trained. So there is a need for feature extraction. The main objective of feature selection is to select a subset of input variables by removing features, which are irrelevant or of no predictive information. Feature selection has proven effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results. Feature selection has a main goal of finding a feature subset that produces higher classification accuracy. Information Gain has been used for this purpose in this paper. Information gain forecasts that whether a word appears or not in a news. Presence or absence of word enables us to select informative features easily. Hence, making classification more durable and reliable.

11. NEWS CLASSIFICATION

After feature selection the next phase is the classification phase which is an important phase in which the aim is to classify the unseen news to their respective categories. In this paper neural network has been used for news classification into various categories. This network is efficient than the other classification algorithm as it is possible to solve the problems using neural networks which were previously insolvable.

12. EXPERIMENTATION AND RESULTS

This section shows the experimental results based on the proposed algorithms. The results for the category classification will be given. The proposed news classifier is designed and developed for classification of news using SVM (Support Vector Machine) and Neural Network.

13. RESULTS

The articles test used a 1000 article subset of BBC news. The subset was made up of several topics. Starting without any information of news topics the news articles were fed into the system in random order. The interested is in how well the articles are grouped together in topics. The evaluation indicators include: Precision, Recall and Accuracy and Elapse Time. The contingency table for results evaluation of classification is shown in table.

- Precision is defined as a fraction of news that is relevant
- Recall is defined as fraction of relevant news that is retrieved.
- True Positive means that news which is classified to its correct class.
- False Negative means that news is classified to a wrong class.
- True Negative means that news which does not belong to that class and is misclassified.
- Accuracy of a news headline is defined as the sum of true negative and true positive.

		PREDICTION CLASSIFICATION	
REAL CLASSIFICATION		P(POSITIVE)	N(NEGATIVE)
	P	TP(TRUE POSITIVE)	FN(FALSE NEGATIVE)
	N	FP(FALSE POSITIVE)	TN(TRUE NEGATIVE)

Recall $R=\frac{TP}{TP+FP}$

Precision $P=\frac{TP}{TP+FN}$

Accuracy $A=\frac{TP+TN}{TP+FN+FP+TN}$

Table 1: Percentage of Accuracy, Precision, Elapse Time, recall values of different news

CATEGORY	**ACCURACY**	**PRECISION**	**ELAPSE TIME**	**RECALL**
SPORTS	0.9129	0.9167	0.267	0.9091
ENTERTAINMENT	0.9354	0.9375	0.265	0.9333
BUSINESS	0.9129	0.9167	0.208	0.9091
HEALTH	0.9309	0.9286	0.140	0.9286

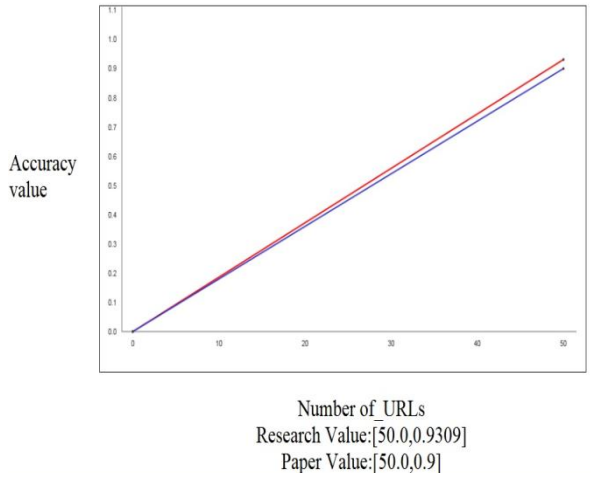

Fig 3: Graph showing the accuracy of proposed method

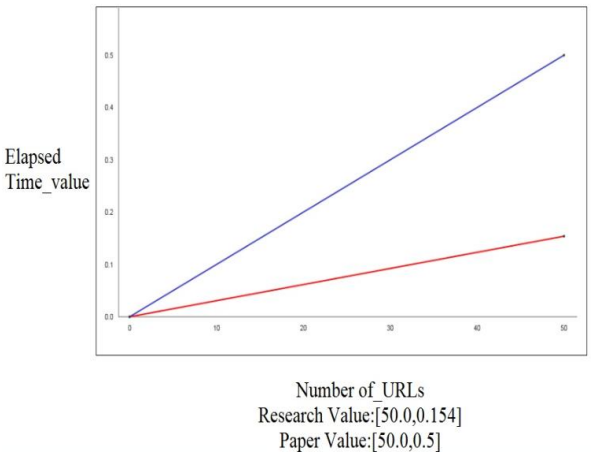

Fig 4: Graph showing the time elapsed using the proposed method

14. CONCLUSION AND FUTURE WORK

In this work, a successful implementation has been done to extract the news from the online portal for the further processing. Along with this the clusters of different categories has been created so that the combination of Support Vector Machines (SVM) and Neural Network along with the WordNet could be applied to it to regain a better efficiency.

In future the work can be implemented to the inner clusters of the existing clusters. Some other algorithm such as Greedy algorithm can be implemented to check if it performs better than neural network. Classification can also be done in other languages such as Urdu. One can also create their own stemmer which can improve classification accuracy.

15. REFERENCES

[1] Deshmukh, R. R., and Mr DK Kirange. J.2013 Classifying News Headlines for Providing User Centered E-Newspaper Using SVM.

[2] Zheng, Hai-Tao, Bo-Yeong Kang, and Hong-Gee Kim. J. 2009 Exploiting noun phrases and semantic relationships for text document clustering.Information Sciences

[3] Li, Jinyan, et al. J.2015 Hierarchical classification in text mining for sentiment analysis of online news.Soft Computing

[4] Rana, Mazhar Iqbal, Shehzad Khalid, and Muhammad Usman Akbar C.2014 News classification based on their headlines: A review.Multi-Topic Conference (INMIC)

[5] Zheng, Hai-Tao, Bo-Yeong Kang, and Hong-Gee Kim.J.2009 Exploiting noun phrases and semantic relationships for text document clustering. Information Sciences

[6] Ramasubramanian, C., and R. Ramya .J.2013 Effective pre-processing activities in text mining using improved porter's stemming algorithm. International Journal of Advanced Research in Computer and Communication Engineering

[7] CHASE, Zach, Nicolas GENAIN, And Orren KARNIOL-TAMBOUR. Learning Multi-Label Topic Classification of News Articles.

[8] Munková, Daša, Michal Munk, and Martin Vozár.J.2015.Data pre-processing evaluation for text mining: transaction/sequence model. Procedia Computer Science

[9] Zhou, Qingqing, and Chengzhi Zhang. J.2014 Sentiment Classification of Chinese Reviews in Different Domain: A Comparative Study.Machine Learning and Cybernetics. Springer Berlin Heidelberg

[10] Bracewell, David B., et al. P.2009 Category classification and topic discovery of japanese and english news articles.Electronic Notes in Theoretical Computer Science

[11] Punitha, S. C., and M. Punithavalli. P.2012 Performance evaluation of semantic based and ontology based text document clustering techniques." Procedia Engineering

[12] Kumar, Rama Bharath, Bangari Shravan Kumar, and Chandragiri Shiva Sai Prasad. J.2012 "Financial news classification using SVM." International Journal of Scientific and Research Publications

[13] Chen, Chun-Ling, Frank SC Tseng, and Tyne Liang. J.2010 .An integration of WordNet and fuzzy association rule mining for multi-label document clustering. Data & Knowledge Engineering

[14] Hassan, Malik Tahir, et al. P.2015 Cdim: Document clustering by discrimination information maximization. Information Sciences.

[15] Li, Huajiao, et al. P.2015 Words Analysis of Online Chinese News Headlines about Trending Events: A Complex Network Perspective. PloS one.

[16] Luo, Congnan, Yanjun Li, and Soon M. Chung. J.2009.Text document clustering based on neighbors.Data & Knowledge Engineering.

[17] Ramdass, Dennis, and Shreyes Seshasai P.2009. Document classification for newspaper articles.

[18] Yan, Yang, Lihui Chen, and William-Chandra Tjhi.J 2013 "Fuzzy semi-supervised co-clustering for text documents." Fuzzy Sets and Systems