Development of Cluster based Supervised Learning Technique for Web News Extraction

Pardeep Kaur
Punjabi University Regional Centre for Information Technology and Management, Mohali

Rekha Bhatia, PhD
Punjabi University Regional Centre for Information Technology and Management, Mohali

ABSTRACT
World Wide Web makes it a prominent source of online information as abundance of data is available on the web and lots of data gets uploaded on daily basis. Due to the presence of massive information on the web it seems easier and simpler to get any information at any time effortlessly, but it requires a lot of focus. Numerous web mining techniques have been studied like extractors, wrappers etc, that provide various methods to extract useful web content. In this paper a semi-supervised web news extraction technique is proposed that uses unsupervised clustering technique and supervised classification technique.

Keywords
Web Mining, Web News, Web News Extraction, Unsupervised Machine Learning, Classification

1. INTRODUCTION
World Wide Web is huge and dynamic source of information online today. It is becoming more popular day by day stepping up the number of its users. The information available on the web is large and it is piling up every second due to its immense popularity. Thus the search for data has become easier, but it is a matter of concern. Search engines return the results on the basis of a query entered; sometimes these results are not relevant or useful. In such situations it becomes difficult to get useful information at right time. To deal with this problem various web mining techniques have been proposed. Web mining is the concept of using several data mining techniques to extract useful information online. It is a converging research area from different research communities such as database, machine learning, natural language processing. Web news extraction is highly researched area of text mining. As there are number of news publications producing heterogeneous content with different formats, makes web news extraction an open challenge. Number of techniques has been adopted to deal with the problem of web data extraction. These techniques are not very popular or efficient, because human intervention is needed in most of the extraction methods and these techniques showed low quality data extraction results. The paper proposes a web news extraction technique using machine learning approach. The proposed technique is combination of unsupervised clustering and supervised classification methods. K-means clustering algorithm is used and SVM classifier is trained on the basis of labelled clusters. Accuracy, recall and precision values are calculated.

2. RELATED WORK
Many different web news extraction techniques have been proposed based on the need of the system. There are many supervised or unsupervised machine learning techniques, DOM tree approaches, template detection, extractors and many more. Feature extraction methods like HRPP (Hyper Sphere-based Relevance Preserving Projection) and ranking function for searching images on one click has been proposed [1]. News extraction techniques like ternary tree approach [2] and DOM tree approaches have been used in various ways. These techniques are used with the combination of machine learning either supervised or unsupervised based on the requirement [3][4]. In DOM tree approaches web page is considered as a tree that consists news blocks as nodes and pre-processing of data is not needed [5]. DOM tree approaches are more time and resource consuming. The recent clustering and classification techniques based on different algorithms such as K-mean clustering algorithm, hierarchical clustering, KNN classification and many more have been used to extract and classify the web content [6]. A similarity measure for text classification that checks the similarity of two documents, this technique is based on textual data only [7]. The unsupervised techniques for unstructured and ungrammatical data that automatically selects the relevant reference sets and uses it for extraction is a good attempt to deal with unstructured web content, but there is a challenge if data is semi-structured [8]. There exist tree edit distance and visual wrapper based techniques. In tree edit distance model the entire web content is extracted and relevant web pages are identified discarding the other content [9]. An online news extraction method based on human perception is used to identify the content. This method simulates the human perception and identification and extracts news data on the basis of format, position and semantics. It performed better than tree edit distance method and visual wrappers.

3. PROPOSED TECHNIQUE
In this section the proposed technique for web news extraction and its working is discussed. First of all web news content is taken and it is cleaned that is called pre-processing of news data. As a result of this the noise (advertisements) from news content is removed. Once the noise is removed the HTML/XML documents are converted into text from. The next steps include clustering and classification of the processed data. Before applying clustering technique the TFIDF values are calculated for every word in the document. The major steps of the applied methodology are discussed below:

3.1 TFIDF (Term Frequency-inverse Document Frequency)
TFIDF is a weight that is used to rank the importance of a term in its contextual document collection. Here TFIDF values are calculated using NLTK library in Python.

3.2 Clustering
K-mean clustering algorithm is used for clustering. Here dataset is a group of text documents taken from News20, so tf-idf values are used as points to make clusters. Once the

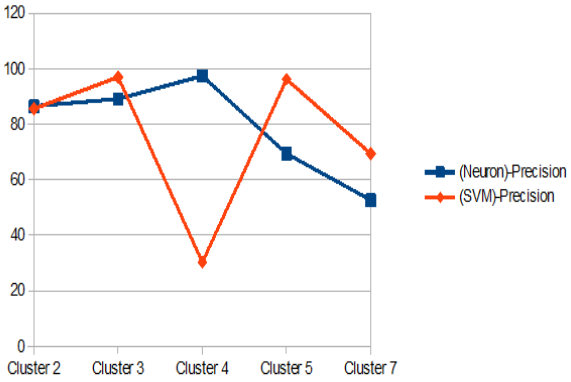
clusters are obtained labelling is done. These labelled clusters are further used in classification.

3.3 Classification

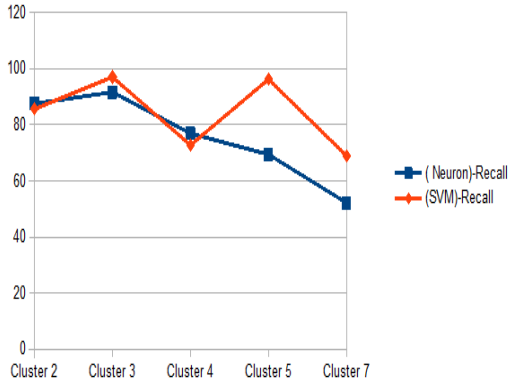
Here SVM (Support Vector Machine) classifier is used for classification. The clusters are selected on the basis of K-mean algorithm and the selected clusters are used for training the SVM. It classifies the data into related classes.

4. RESULTS

The proposed technique is implemented on MATLAB after clustering process. Results have been built on the basis of confusion matrix generated by SVM and then recall, precision, accuracy, and F1-Measure is calculated. Confusion matrix is also known as an error matrix. Each column represents the instances in a predicted class while rows represent instances of actual class. Accuracy, precision and recall are computed after training the SVM in MATLAB. Values are estimated taking different number of classes in a confusion matrix. Minimum number of classes taken is 2 and maximum number of classes used is 6. The results obtained with SVM are compared with Neural Network classifier and it showed that on average SVM's performance is better than NN. The comparison graphs between SVM and NN are given below:



Precision comparison graph between SVM and NN



Recall comparison graph between SVM and NN

5. CONCLUSION AND FUTURE WORK

The suggested method is for extracting news from the internet web pages by using the idea of clustering with neural genetic approach. Extracting web news content is some other process for examining a pattern in the web information that is communicated as regular appearance in the contribution of web pages. We proposed a method for pre-processing the web content using text mining and classified it using a cluster based supervised learning approach. The complexity is decreased and accuracy is improved of web news extraction by this technique. Local noise can be discovered and removed that gives web news in a proper format. Thus the extraction process is enhanced. The proposed approach is analysed by precision, recall and accuracy. Future work involves experimentation of the proposed method with different data sets and optimization of the proposed technique. Here we used web news as input data that is of structured form majorly, so in future data of different format can be used like unstructured or semi-structured.

6. REFERENCES

[1] Zhong Ji, Member, Yanwei Pang, Senior Member, and Xuelong Li, "Relevance Preserving Projection and Ranking for Web Image Search Reranking", VOL. 24, NO. 11, NOVEMBER 2015.

[2] Debina Laishram and Merin Sebastian, "Extraction of web news from web pages using a ternary tree approach," IEEE Second International Conference on Advances in Computing and Communication Engineering,, pp. 628-633, 2015.

[3] Shanchan Wu, Jerry Liu, Jian Fan, "Automatic Web Content Extraction by Combination of Learning and Grouping," International World Wide Web Conference Committee (IW3C2), pp. 1264-1274, WWW 2015, May 18-22, 2015, Florence, Italy.

[4] Yan Guo et al, "ECON: An Approach to Extract Content from Web News Page," IEEE 12th International Asia-Pacific Web Conference, 2010, pp. 314-320

[5] Yongquan Dong1,Qingzhon Li1,Zhongmin Yan1 and Yanhui Ding," A Generic Web News Extraction Approach," Proceedings of the 2008 IEEE, International Conference on Information and Automatio, Zhangjiajie, China,June 20-23,2008.

[6] M. Wook, Y. H. Yahaya, N. Wahab, M. R. M. Isa, N. F. Awang, and H. Y. Seong, (2009) "Predicting NDUM student"s academic performance using data mining techniques," in Proc. 2009 Second Int. Conf. Comput. Electr. Eng., pp. 357-361.

[7] Yung-Shen Lin et al, "A Similarity Measure for Text Classification and Clustering," IEEE transactions on knowledge and data engineering, vol. 26, no. 7, pp. 1575-1590, July 2014.

[8] Matthew Michelson and Craig A. Knoblock, "Unsupervised Information Extraction from Unstructured,Ungrammatical Data Sources on the World Wide Web," International Journal of Document Analysis and Recognition (IJDAR), August 2007.

[9] Davi de Castro Reis et al, WWW2004, New York, USA.ACM1-58113-844-X/04/0005. "Automatic Web News Extraction Using Tree Edit Distance,"May 17.22, 2004, pp. 502-511