Classify News Articles Based On Location Using Machine Learning

Prof Minal P. Nerkar, Mayuresh Zende, Ganesh Rupnawar , Bhujang Zanak, Kanchan Ghumre

Department of Computer Engineering
AISSMS Institute Of Information Technology, Pune, India

Abstract: With the influx of myriad news accompanied with busy lifestyle, there is a pressing need to classify news according to the requirements of an individual. People are generally more interested what is going on, in their immediate surroundings. In this report, we model this problem by classifying the news articles based on cities and providing the entity with the collection of city specific news. We have developed our own server for content extraction from the HTML pages of news articles. **PMML Decision making Algorithm** have been employed and the accuracy has been noted. Results exhibit that machine learning techniques can be harnessed to achieve our goal and thus calls for further research to improve the efficiency of solving this issue.

I. INTRODUCTION

Most of the news articles, though informative, might be of less relevance to an individual. Hence, it poses a mammoth task for extracting relevant news with respect to an individual. The interests can depend on several factors like type of the news articles, place to which the news belongs to, etc. In this case, we have considered the interest based on geographical domain. For example, a person wants to read news specific to Mumbai and is provided with a flood of news relating to all the cities of India. In this case, it would be cumbersome for the person to and the city specific news. In this system, we have implemented machine learning techniques to classify news articles belonging to a particular location. The location can be a city, state, country, etc but we have examined the results based on cities.

In this system, we describe an algorithm devised to be used for the summarization of long text, which will be based on semantic of the sentences. It retrieves the most important sentences from a text by the most important keywords and these keywords also found by automatically. We can perform this process either in basic or compound mode.

II. MOTIVATION

News plays a very important role on daily basis. In Traditional way reading a news was a time consuming and hectic work. That's why our goal is to create an android application so that user get news as per their interest and preferred location in summarized form.

III. PROBLEM STATEMENT

To develop an Android app so that the user will be able to fetch classified and categorized news based on his current as well as specified location using Machine Learning.

IV. ARCHITECTURE DIAGRAM

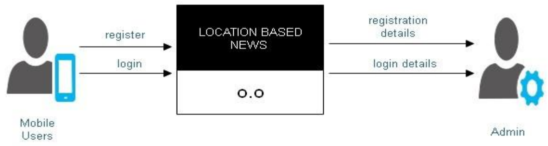

Fig 1. User Registration.

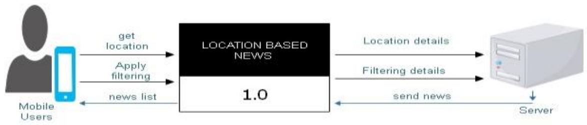

Fig 2. User Requesting News.

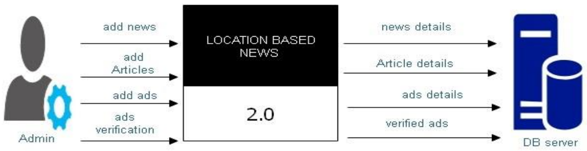

Fig 3. Admin Portal

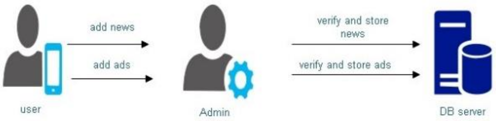

Fig 4. User – Admin Interaction

V. RELATED WORK

The important step of research is to Literature survey first Before where we actually start developing our system, we need to study the previous papers which should be of our domain which we are working on. The basis of study is that we can generate the drawback or predict. Then start working with the reference of previous papers.
Here, we review the related work briefly on News.

Reference paper [1] With the influx of myriad (dense) news accompanied with busy lifestyle, there is a pressing need to classify news according to the requirements of an individual. People are generally more interested what is going on, in their immediate surroundings. In this paper, we model this problem by classifying the news articles based on cities.

Reference paper [2] Stemming and Lemmatization are two significant natural language processing techniques extensively used in Information Retrieval for query processing and Machine.Translation for reducing the data sparseness.

Reference paper [3] Support Vector Machines (SVMs) have been extensively researched in the data mining and machine learning communities for the last decade and actively applied to applications in various domains.

Reference paper [4] Traditionally, machine learning algorithms have been evaluated in applications where assumptions can be reliably made about class priors and/or misclassification costs. In this paper, we consider the case of imprecise environments, where little may be known about these factors and they may well vary significantly when the system is applied

Reference paper [5] In many real-world scenarios, the ability to automatically classify documents into a fixed set of categories is highly desirable. Common scenarios include classifying a large amount of unclassified archival documents such as newspaper articles, legal records and academic papers.

Reference paper [6] The navie bayes classifier, currently experiencing a renaissance in machine learning. Has long been a core technique in information retrieval. We review some of the variations of navie Bayes model used of text retrieval and classification .

Reference paper [7] Random forests have proved to be very effective classifiers, which can achieve very high accuracies. Although a number of papers have discussed the use of fuzzy sets for coping with uncertain data in decision tree learning, fuzzy random forests have not been particularly investigated in the fuzzy community

Reference paper [8] Web-page classification is much more difficult than pure-text classification due to a large variety of noisy information embedded in Web pages. In this paper, we propose a new Webpage classification algorithm based on Web summarization for improving the accuracy.

Reference paper [9] Support Vector Machines (SVMs) have been extensively researched in the data mining and machine learning communities for the last decade and actively applied to applications in various domains.

VI. PROCESSING STEPS

PHP Semantic science:

Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. Automatic data summarization is part of machine learning and data mining. The main idea of summarization is to find a representative subset of the data, which contains the information of the entire set. Summarization technologies are used in a large number of sectors in industry today.

A) The algorithm of this implementation is:

1) Find sentences

2) Remove stop words

3) Create integer values by find and count the matching words

4) Change the integer values by the related word's integer values

5) Normalize values to create scores

6) Order by scores

B) Web-Services:

Web Service is can be defined by following ways:

- is a client server application or application component for communication.
- method of communication between two devices over network.
- is a software system for interoperable machine to machine communication.

In our application we are using PHP services and JSON for communication.

VII. CONCLUSION

we have determined the possibility of using machine learning algorithms to classify the news articles based on location. The experiments show that this problem can be successfully solved by using various Classifiers such as PMML. PMML allows you to easily share predictive analytic models between different applications. It provides a greater readability to the user that’s why it is used over other classifiers. We can also specify biasing nature for particular node if needed.

The proposed system can be used as a part of more complex news article classification systems. Our future goal is to improve the accuracy and also try classifier like Neural Network. We can further increase the number of the input articles to 1000 folds compared to our present dataset for training our model in order to improve on our results.

REFERENCES

1) Jayant sachdev, Vignesh Rao, A Machine learning approach to classify news article based on location, 2017
2) M.Kasthuri, Dr.S.Britto Ramesh Kumar, A Framework for Language Independent Stemmer Using Dynamic Programming. , 2015
3) H. a. K. S. Yu, SVM tutorial: Classification, regression, and ranking, 2009.
4) T. Landgrebe, P. Paclk, R. Duin, and A. Bradley, Precision-recall operating characteristic (P-ROC) curves in imprecise environments, 2006.
5) C. Ee and P. Lim, Automated online news classification with personalization.
6) D. Lewis, Naive (bayes) at forty: The independence assumption in information retrieval, 1998.
7) J. Davis and M. Goadrich, The relationship between precision recall and ROC curves, in Proceedings of the 23rd International Conference on Machine Learning, ser. ICML 06. New York, NY, USA: ACM, 2006, pp. 233240 [Online].Available:http://doi.acm.org/10.1145/1143844.1143874
8) M.Kasthuri, Dr.S.Britto Ramesh Kumar, A Framework for Language Independent Stemmer Using Dynamic Programming, International Journal of Applied Engineering Research, ISSN 0973- 4562 Vol.10, pp 39000-39004, Number.18, 2015.
9) H. a. K. S. Yu, SVM tutorial: Classification, regression, and ranking, Handbook of Natural Computing 2009.