ISSN 2395-1621

Infology Categorization by Data Acquisition

[#1]Tanmay Sardesai, [#2]Dheeraj Shetty, [#3]JaykumarBhati, [#4]Kiran Sabale.
[#5]Prof. Sandeep Gore.

[1]tanmaysardesai136@gmail.com
[2]dheerajshetty5010@gmail.com
[3]jaykumar1994.jk@gmail.com
[4]kiransaabale93@gmail.com

[#1234]Department of Computer Engineering
[#5]Prof. Department of Computer Engineering

G.H.Raisoni College of Engineering and Management
Pune, Maharashtra, India.

ABSTRACT

The popularity of Internet Services is increasing in the recent years, especially for accessing all the required information in our day-to-day lives. Web Browsers are used for this purpose. Users sometimes get some information which is not being intended to be shown to their particular age group. We propose to use an enhanced system for file accessing considering security, ease of access, performance and few other parameters. In this project, we use text mining technique to categorize the data according to age group and also user interest.Web content classification using machine learning techniques is therefore an emerging possibility to automatically maintain services for the web. The concept of Naïve Bayes classifier is then used on derived features and finally proposed algorithm has been implemented and tested.

ARTICLE INFO

Article History

Received :24th January 2016

Received in revised form :

24th January 2016

Accepted :27th January , 2016

Published online :

27th January 2016

***Keywords*: Text mining, categorize data, Web content classification, Naïve Bayes classifier**.

I. INTRODUCTION

In the modern world, computers have become inevitable in all the fields where precise planning, analysis, calculations are needed. Computers have become part and parcel of day to day life. They are used in various fields like industries, Banks, Railways, Business centers, Educational Institution, E-Market, Offices etc. The 21st century is provided with many technological innovations still we don't have proper management of contents in various websites. Our project helps to take the idea further. The main aim of our project is to categorize the contents according to the user's age. With the explosive growth of the textual information from the electronic documents and World Wide Web, proper classification of such enormous amount of information into our needs is a critical step towards the business success. However, it is time-consuming and labor intensive for a human to read over and correctly categorize an article manually. Attempts to address this challenge, automatic document classification studies are gaining more interests in text mining research recently. Consequently, an increasing number of approaches have been developed for accomplishing such purpose, including k-nearest-neighbor (KNN) classification ,Naïve Bayes classification, Support Vector Machines (SVM), Decision Tree (DT), Neural Network (NN), and maximum entropy. Among these approaches, the Naïve Bayes text classifier has been widely used because of its simplicity in both the training and classifying stage. Thus, the present study focuses on employing Naïve Bayes approach as the text classifier for document classification and thus evaluates its classification performance against other classifiers.

II. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language can be used for developing the tool.

A Survey of Naive Bayes Machine Learning approach in Text Document Classification Year: 2010

In this paper, we have studied that Naive Bayes works well in large datasets even with the simple learning algorithm had been a great inspirations in doing this survey. The Naive Bayes technique performs better and yields higher classification accuracy when combined with the other techniques.

Text Classification Using Data Mining Year: 2005

In this paper, we have studied that the existing techniques require more data for training as well as the computational time of these techniques is also large.In contrast to the existing algorithms, the proposed hybrid algorithm requires less training data and less computational time.

III. PROPOSED METHODOLOGY

The proposed method to classify text is an implementation of Naïve Bayes Algorithm. In this first collect the large data items on the electronic form. Then after remove the noise by using data cleaning techniques. Now implement the Naïve Bayes Algorithm and to find out the keywords of the data for all category related topics and obtain probability using Naive Bayes Classifier.

ALGORITHM:

The algorithm that is being used is Naïve bayes Classifier Algorithm. This algorithm is as follows:

Likelihood Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

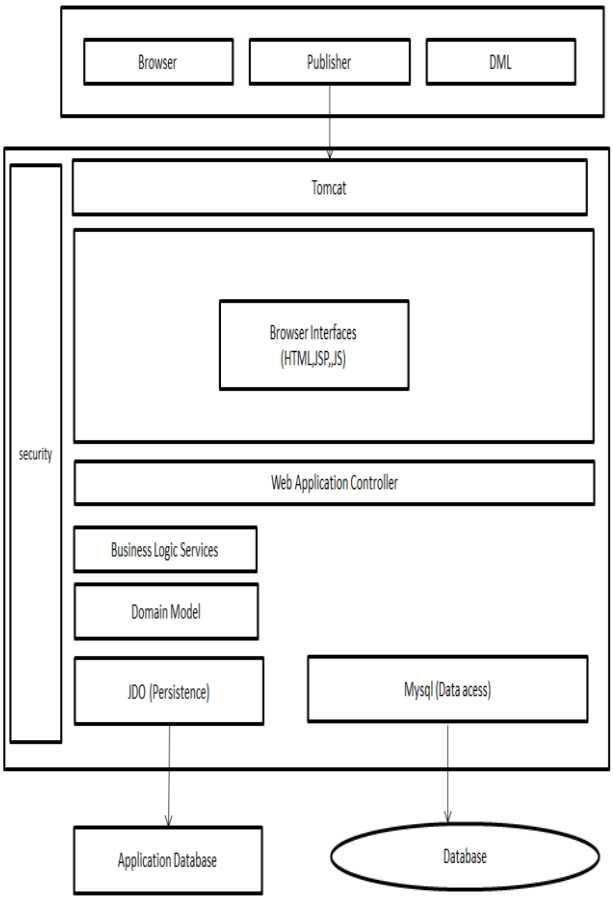
Posterior Probability Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Above,

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*

ARCHITECTURE



The System architecture consists of 3 tiers

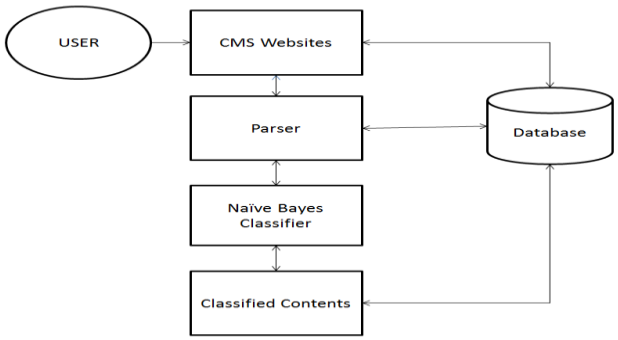
(1) Client tier
(2) Middle tier
(3) Back-end tier.

- Users insert the queries from the **Client tier** or the front end. Front end is basically GUI between the system and the users. Front end consists of Browser which is interface to user like html pages. Publisher is a block in front end which helps the system to show the data in graphical format, while DML i.e. Data Manipulation Language is used for registering or login of users.

- **Middle tier** consists of Tomcat which is a web server used to store all the webpages other than the data. It contains MVC model, where m is model, v isview, c is controller. Model is Jsp, view is Html pages and controller is servelet which contains .java files.

- **Back-end** tier contains the database where all the data is stored in database like MySQL.

IV. WORKING



When the user enters the queries, the content is sent to CMS i.e. Content Management System which breaks the data down in tokens by parsing it. Thesetokens are classified into different datasets by naive bayes classifier and thenthis classified content is stored in database.

V. CONCLUSION

Text Document Classification has been in research for decade inwhich various researchers has experimented with available machine learning techniques in which each method has been aimed to improve the classification accuracy; Among which Naive Bayes works well in large datasets even with the simple learning algorithm had been a great inspirations in doing this survey.From the survey the inference made is that the Naive Bayes technique performs better and yields higher classification accuracy when combined with the other techniques. The other inference is that Multinomial Naive Bayes event model is more suitable when the dataset is large when compared to the Multi-variant Bernoulli Naive Bayes Model. This paper presented an efficient technique for text classification. The existing techniques require more data for training as well as the computational time of these techniques is also large. In contrast to the existing algorithms, the proposed hybrid algorithm requires less training data and less computational time. In spite of the randomly chosen training set we achieved encouraging, it would be better if we work with larger data sets with more classes.

REFRENCES

1. Text Classification Using Data Mining, ICTM 2005, Published by S. M. Kamruzzaman,Farhana Haider, Ahmed Ryadh Hasan.

2. A Survey of Naive Bayes Machine Learning approach in Text Document Classification, (IJCSIS) International Journal of Computer Science and Information Security, Vol.7, No. 2, 2010, Published by Vidhya.K.A, G.Aghila.

3. Is Naive Bayes a Good Classifier for Document Classification, International Journal of Software Engineering and Its Applications Vol. 5, No. 3,

   July, 2011, Published by S.L. Ting, W.H. Ip, Albert H.C. Tsang.

4. Automated Classification of Web Sites using Naive Bayesian Algorithm, IMECS2012, March 14-16, 2012, Published by Ajay S. Patil, B.V. Pawar.

5. Enhanced Classification Accuracy on Naive Bayes Data Mining Models, International Journal of Computer Applications (0975 {8887) Volume 28{ No.3, August 2011, Alamgir Hossain, Keshav Dahal, Md.Faisal Kabir.

