IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 21-25
www.iosrjournals.org

Content Systemization using Naive Bayes Technique

Isha Pandya, Deepti Theng
Computer Science and Engineering G. H. Raisoni College of Engineering Nagpur, India
Computer Science and Engineering G. H. Raisoni College of Engineering Nagpur, India

Abstract- *In today's scenario, the internet is growing day-by-day. And so the data and information is growing large. More and more users rely on search engines. It is the need of the hour to improve the speed and efficiency of the search engine. When a user searches a query he should be able to get most relevant results in very less time. Therefore there is need to do content systemization for efficient and faster search. One of the efficient ways to achieve is to use a supervised learning algorithm called Naive Bayes algorithm.*
Keywords- *Content systemization, supervised learning, Naive Bayes Technique*

I. Introduction

World Wide Web is growing day by day and so the web pages are growing from thousands to millions. When user is searching for a query in search engine, he is expecting the correct and most relevant result. Because of this reason Content systemization has become important.

Content systemization is the process to group documents according to some predefined knowledge. Documents of same concept are grouped together. And non-related documents are in other category. For grouping of documents clustering techniques will be used.

Machine learning is an important concept used here. Machine learning is programming computers to optimize a performance criterion using example data or past experience. There are two things that need to be achieved in the learning process. First, the training needs to be done with an efficient algorithm to solve the optimized problem, and to store and process the data. Second, the trained model needs an efficient representation and algorithm for inference.

Naive Bayes technique is a supervised learning algorithm. These algorithms generate a function based on the training data set that maps inputs to predetermined labels or classes.

II. Literature Survey

There are many algorithm like K-Nearest Neighbor, Neural networks, Support Vector Machine, decision tree. Many researchers have used different techniques for systemization. According to S.L. Ting.et.el, Naive Bayes is the best classifier used for content systemization. NB gives more accurate results than Support Vector Machine, NN and decision tree. NB gave 97% accuracy, SVM gave 96.6% accuracy, NN gave 93% accuracy, and Decision tree gave 91.1% accuracy. Time required is less as compared to other four algorithms. K-Nearest Neighbor is a lazy classifier and is bad if speed is important Naïve Bayes is faster than K-NN. NB handles missing data while k-NN cannot.

According to Daniela Xhemali.et.el,[8] Acuracy of Naïve bayes classifier is 95.20% while that of Decision tree is 94.85%.Precision of Naïve bayes classifier is 99.37% while that of decision tree is 98.31%. Recall of Naïve bayes classifier is 95.23% while that of decision tree is 95.90%. F-Measure of NB classifier is 97.26% while that of decision tree is 97.09%. Tokenization with preprocessing gives more accurate and fast result as compared to tokenization without preprocessing.

III. Proposed system

The input of the system is URL. The contents of the URL is classified into predefined groups with the help of training data.
URL contents:

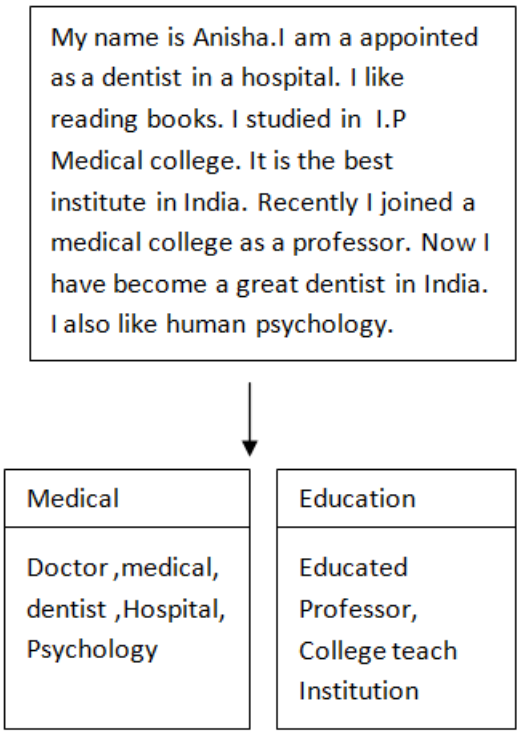

Fig. 1 Predefined categories

IV. Methodology

The input is supplied as a URL. The Contents of URL will undergo through following steps.

The first step is preprocessing. It includes Tokenization. Given a sequence of string as input, tokenization is the task of chopping the string into pieces called Tokens. It removes some characters like punctuation. Second step in preprocessing is removal of stop words. Some stop words are defined. And if in the contents of URL these words occur, they are removed because they have no significance in content systemization. Third step is to count frequency of words and sort them. Then preprocessed and tokenized text is obtained. This text is given as input to classifier. The classifier takes URL and output of training module as input and classifies it into a category. It uses Naïve Bayes theorem is based on concept of probability

1.Prior probability (A) = (Probability of URL occuring in class)

no. of words in A/total no. of words

2.Likelihood of X (new) given A =

no. of words in A in the vicinity of X/total no.of A

3.Posterior probability of X being in A,

log(Prior probability) +log(likelihood of X being A)

where A= category and Highest probability value is the winner.

The intermediate step is to train the classifier. In training the classifier some categories are defined which contains some pre-defined words and serializes them to make learning data. These categories are given as input to trainer module. The trainer module finds the probability of each word. It serializes them and stores it in wordprob.srl which acts as local database. The output of training module is word probability.

V. Result and discussion

The URL of a website is taken as input.

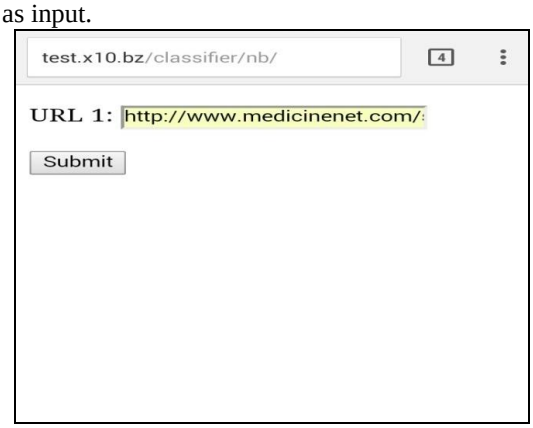

Fig. 2 Request URL for Searching

Tokenization is done to the contents of the website and removing punctuation marks are removed.

Fig. 3 Tokenization

Stop words are removed from the tokenized text.

Fig. 4 Removal of stop words

The words are sorted in the decreasing order of their frequency of occurrence.

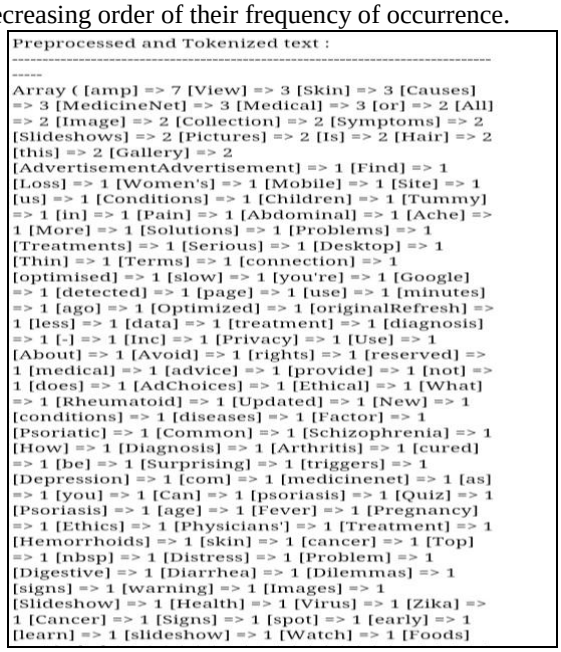

Fig 5. Preprocessed and tokenized text

Preprocessed and tokenized text is given input to the classifier.

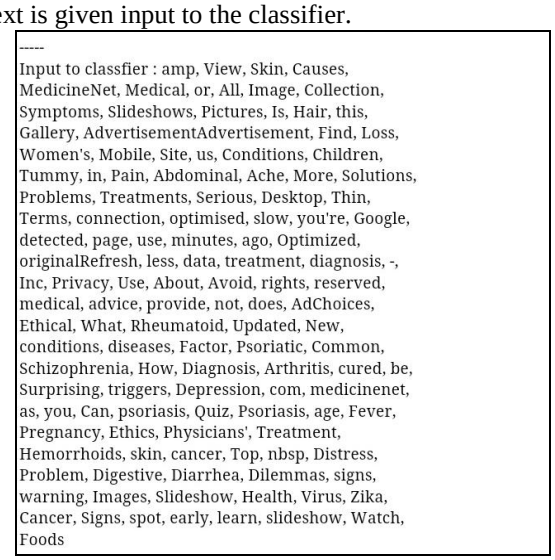

Fig 6. Input to classifier

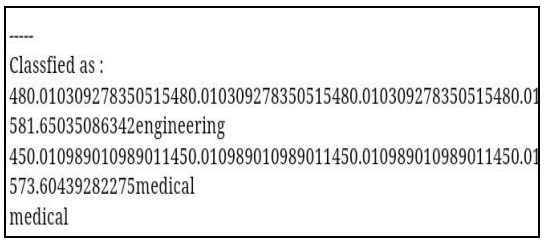

Fig. 7 Classified output

VI. Future work

To further improve the search results and quality of classification we can apply Feature selection method. It makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. Furthermore preprocessing techniques like pruning, stemming can be applied in step 1 for

improving quality of search result. Semantic analysis can also be considered for more accurate classification. Semantic analysis refers to formal analysis of the meaning.

VII. Conclusion

We succeeded in building NB classifier Tokenization with Preprocessing techniques are applied to improve the quality of systemization and search quality. Supervised learning technique called Naïve Bayes is applied for content systemization. The disadvantages of SVM, Neural network, Decision tree and K-nearest neighbor are overcome by Naïve bayes technique. The contents of URL obtained dynamically are classified accurately by this classifier. This gives high performance of search engines. We also presented some preprocessing techniques like tokenization, removal of stop words and frequency count of words. Preprocessing has huge impact on the performance of systemization. It improved quality of classification.

References

[1]. Muhammad Husnain Zafar1 and Muhammad Ilyas2 ," A Clustering Based Study of Classification Algorithms ," International Journal of Database Theory and Application Vol.8, No.1 (2015)

[2]. Soumen Swarnakar1, Sangita Karmakar2 , " CONCEPT BASED CATEGORIZATION OF DOCUMENTS FOR SEARCH ENGINES "Volume: 04 Issue: 10 | Oct-2015

[3]. Teena Rani, Ankush Goyal ,"Survey of Clustering Techniques for Information Retrieval in Data Mining ", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 4, April 2015

[4]. Khushboo Thakkar, Urmila Shrawankar, "Test Model for Text Categorization and Text Summarization" , International Journal on Computer Science and Engineering (IJCSE) ISSN :0975-3397 Vol. 3 No. 4 Apr 2011.

[5]. Bhumika1, Prof Sukhjit Singh Sehra2, Prof Anand Nayyar3, "A REVIEW PAPER ON ALGORITHMS USED FOR TEXT CLASSIFICATION", *International Journal of Application or Innovation in Engineering & Management (IJAIEM),* Volume 2, Issue 3, March 2013.

[6]. JIANG Xiao-Yu, FAN Xiao-Zhong, Wang Zhi-Fei,Jia Ke-Liang, "Improving the Performance of Text Categorization using Automa,tic Summarization", 978-0-7695-3562-3/09 $25.00 © 2009 IEEE DOI 10.1109/ICCMS.2009.29

[7]. S. Revathi , Dr.T.Nalini ," Performance Comparison of Various Clustering Algorithm ", International Journal of Advanced Research in Computer Science and Software Engineering , © 2013, IJARCSSE .

[8]. Daniela XHEMALI, Christopher J. HINDE and Roger G. STONE," Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages", IJCSI International Journal of Computer Science Issues, Vol. 4, No. 1, 2009

[9]. S.L. Ting, W.H. Ip, Albert H.C. Tsang Is Naïve Bayes a Good Classifier for Document Classification International Journal of Software Engineering and Its Applications Vol. 5, No. 3, July, 2011