ISSN: 2321-7782 (Online)
Volume 3, Issue 4, April 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Web Page Classification Using Apriori Algorithm and Naïve Bayes Classifier

Sneha K. Dehankar[1]
Department of Computer Science & Engineering,
Government College of Engineering,
Amravati – India

K. P. Wagh[2]
Department of Computer Science & Engineering,
Government College of Engineering,
Amravati – India

Dr. P. N. Chatur[3]
Department of Computer Science & Engineering,
Government College of Engineering,
Amravati – India

Abstract: Classification is the process of dividing the data into number of groups which are either dependent or independent of each other and each group acts as a class. The task of classification can be done by using several methods by different types of classifiers. The main purpose of the paper is to analyze the task of classification of web page into multiple classes and to learn that how to achieve high classification accuracy in classifying the pages. The classes are predefined in which the web page is to be classified. Web pages are extracted by submitting the keyword to the search engine and then preprocessed to get only relevant words. The Apriori algorithm is used to get the links that are having the associated word sets which are having the value more than the minimum threshold defined. From these the probability of each word will be calculated and page will be classified into its predefined class based on the highest posterior probability calculated. The Naïve Bayes classifier is used to calculate the probability of each word with respect to each class. The proposed web page classification system shows the F-measure value of 75.91%. An analysis of the system reveals that the proposed classification method works well even when the number of web pages is increased. The proposed classification method is suitable for document classification and provides better result as compared to the simple implementation of Naïve Bayes classifier.

Keywords: Apriori algorithm, classification, Naïve Bayes classifier, text categorization.

I. INTRODUCTION

Data mining is the field used to get the relevant data from the huge amount of available data. Many algorithms are used to divide the data into many sections to get the useful information from those respective categories. Methods used for dividing such a huge amount of data are classification, clustering, ranking etc. In this paper classification method is focused for getting the relevant and specific kind of data related to our search. The classification technique divides the information into number of classes to get the data related to that class only. In this work the classification is focused on classifying the web pages into multiple categories based on the content. Document classification is one of the types of text categorization. In the task of document classification when more than two classes exists then it can be termed as multi-class document classification. In case of multi-class document classification when a document belongs to more than one label then it can be termed as multi-label document classification. In web page classification, the all HTML tags, anchor tags etc. are removed. The content of the web page is considered after removing the all tags contained in the web page. The results are extracted for the keyword searched and on that result the algorithm is applied to get the feature words. In Apriori algorithm, output is the frequent item-sets. From these frequent item-sets the correlation is made. Association rule is applied on the frequent item-sets to find the association between the item-sets. The number of occurrences of the words is counted and the association between all the frequent words is determined. Based on these associated word set, the probability for each word is calculated and on the basis of that probability the document is classified into its respective class. If the support value is increased then the classification accuracy is increased. Increasing support value reduces complexity and time required to access whole data. Thus there remains no scanning of whole database using algorithm. The benefit of apriori algorithm is used in this which reduces the whole links to be scanned for getting the relevant links related to the keyword.

In the web page classification approach the classes are first predefined then according to the user search, the links are retrieved from search engine. Stemming is applied on that results extracted and stop words are removed from it to get only useful words from the web pages. Apriori algorithm is used to get only relevant and most related links related to the keyword submitted by the user. Among those extracted links by using algorithm, the links are classified into the classes based on the keyword matched with that respective class. The links are classified into the classes on the basis of their probability. The Naïve Bayes is used to calculate probability of each word of each document. On this probability the links are classified into their respective category.

In many methods of classifying the web page, only tags and URLs are considered while classifying the web page. In this paper the text of the web page is also considered to classify it on the basis of its context. The context based classification shows more accurate results than classifying the page on the basis of tags and URLs.

II. Related Work

Ajay S. Patil, B.V. Pawar [1], proposed the NB approach for classification of home pages for the ten categories considered. This approach can be used by search engines for effective categorization of websites to build an automated website directory based on type of organization.

Igor Kotenko, Andrey Chechulin, Andrey Shorov, and Dmitry Komashinsky [2], considered the problem of automated categorization of web sites for systems used to block web pages that contain inappropriate content. In the paper they proposed the techniques of analyzing the html tags, tags, URL addresses and other information using Data Mining methods and machine Learning.

Shiju Sathyadevan, Athira U, Sarath P R, Anjana V [3], proposed methods for classification of documents using Naïve Bayes classifier. Naive-Bayes algorithm proceeds by finding out the feature vector associated with each category. On the basis of this probability calculated, the documents are classified into their respective category. Based on the precision and recall, the accuracy of all methods are compared based on the time required for training and classification of documents. Result is obtained with the consideration of context of page.

Abu Nowshed Chy, Md. Hanif Seddiqui, Sowmitra Das [8], proposed an approach that provides a user to find out news articles which are related to a specific classification. They used their own developed web crawler to extract useful text from HTML pages of news article contents to construct a Full-Text-RSS. Each news article contents are tokenized with a modified light-weight Bangla Stemmer. In order to achieve better classification result, we remove the less significant words i.e. stop word from the document. They apply the naive Bayes classifier for classification of Bangla news article contents based on news code of IPTC. In this paper they present a system of automatically classifying Bangla News documents. This system provides users with efficient and reliable access to classified news from different sources.

III. Proposed Work

The method for classification of web pages is consisting of applying Apriori algorithm to the results extracted after submitting the query to the search engine. Once the frequent words are taken out then Naïve Bayes is used to calculate probability of each feature to classify the page into the respective class based on its probability. The number of classes has to be predefined in order to classify the pages into the categories. In this total 9 categories are predefined to classify the web pages into those categories. Hotel, Hospital, Computer, Sports, Academic Institutions, Bank, Tours & Travel, Domestic Applications, Automobiles and Other class. The other class label is used to classify the web pages that are not related to the all 9 classes that are mentioned above. Figure 1 shows the architecture of web page classification system to classify the web pages into multiple classes that are predefined for the system.

The steps of web page classification system are defined as:

A. Extracting the results

Keyword is submitted to the search engine to get the results from it. The numbers of links are extracted after submitting the keyword. Among that all links extracted, the classification is applied to get only relevant results related to the keyword submitted.

B. Preprocessing

The links extracted after applying Apriori algorithm are then preprocessed. The stop words are removed first then stemming algorithm is used to get the stemmed words from each web page. Based on these stemmed words obtained, the probability for each word is obtained to classify them into categories. The porter stemming algorithm is used to remove all unnecessary words of web pages.

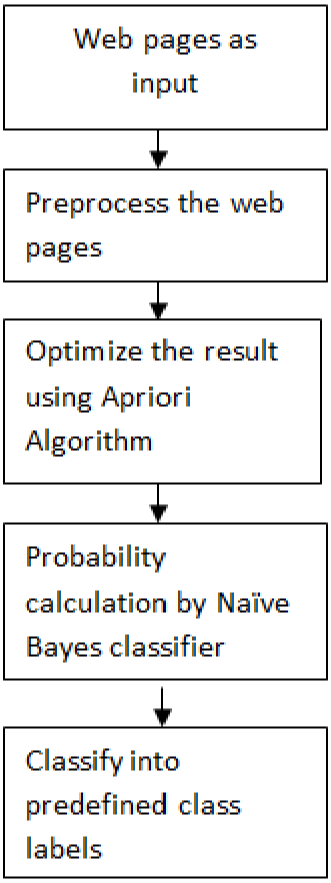
Web pages as input
Preprocess the web pages
Optimize the result using Apriori Algorithm
Probability calculation by Naïve Bayes classifier
Classify into predefined class labels

Fig.1 System Architecture of web Page Classification

C. Apriori Algorithm

The Apriori algorithm works as:

1. Determine the frequent itemsets from each web page using the minimum support value.

2. Apply association rule to the frequent items obtained from step 1 to get the frequent 2- itemsets.

3. Calculate frequent 3- itemset from the result obtained from step 2.

4. Count the number of times associated word sets that are obtained from step 3.

5. If the count >minimum support value then save it else discard.

In the web page classification system, the results are optimized using the minimum support value defined and the words that are having occurrences more than the minimum support value defined, are taken as frequent item-sets else they are discarded. So only that links are taken out as Apriori results which are having associated word set higher than the threshold value defined as minimum support for reducing the number of links needed to classify. It saves the time to classify all the links that are irrelevant to our search. In this the threshold of 0.6 is taken for the first iteration then for getting the frequent 2-itemsets, threshold is taken as 0.3 to optimize the result. The web pages those satisfy the predefined threshold value will be classified only by using the Naïve Bayes classifier else are discarded.

D. Naive Bayes Classifier

After getting the optimized links by applying Apriori algorithm, the posterior probability of each word of each web page is calculated w.r.t. each class label.

The Naïve Bayes Classifier works as:

1. Calculate the prior probability of each class. In this classification system there are 9 categories, so the prior probability for each class is 0.9.

2. Calculate the word likelihood for each word of each web page w.r.t. each class.

3. Determine the posterior probability from the prior probability and word likelihood obtained from steps 1 and 2 respectively.

Posterior probability is calculated as:

$$\mathrm{P}(\frac{C}{D})=\frac{P(C)P(\frac{D}{C})}{P(D)} \qquad (1)$$

P(C) is the prior probability of class w.r.t. document. It is determined as:

$$\mathrm{P(C)}=\frac{Nc}{N} \qquad (2)$$

Where $\mathrm{N_c}$ -Number of occurrences of the documents in class c

N-total number of documents

p(D/C) is the word likelihood calculated as:

$$\mathrm{P}(\frac{D}{C})=\mathrm{P}(\frac{W}{C}) \qquad (3)$$

Where p(w/c)-probability of word w.r.t. class

For calculating the word likelihood, the presence of each word in each web page for particular class label is checked. If the word is present in the predefined class for a particular web page then its feature count is taken as 1 else 0.Similarly, presence of that word into other classes is also checked. To avoid the zero probability error the weight is assume as 1 to solve this problem.

Total posterior probability is then calculated as:

$$P\left(\frac{C}{D}\right) = P(C)P(\frac{W}{C}) \qquad (4)$$

E. Classification into predefined class labels

The link of web page is classified into the predefined class to which it belongs. The predefined word sets for each class label is taken into the consideration while classifying the web pages. The number of matching words from each class label is checked w.r.t. each web page. It is done on the basis of highest probability calculated for the word w.r.t. each class label. The probability of each class is checked for each word and the web page is classified into the class that is having higher probability as compared to all other classes.

IV. Experimental Results and Discussion

The various keywords are submitted to the search engine and the respective values of precision and recall for that keywords are calculated as shown in Figure 2. From these values then F-Measure value is calculated. In the work of classification, i.e. "Web page classification using data mining" [4], by Keyur Patel and Ketan Sarvakar proposed the system to classify the web pages into predefined classes using Apriori algorithm and Naïve Bayes classifier by working on training dataset. They checked the accuracy up to 55% of dataset. Their classification system has shown better accuracy when 50% of dataset was used and when increased the dataset percentage then the accuracy was decreased. So the average accuracy obtained from their proposed system is 59.76% while our proposed work is giving the F-measure value of 75.91%.From this we are getting the accuracy of 61.98%.This shows that the accuracy of the proposed work is improved. The precision and recall obtained by submitting the various keywords to the system shows that the number of correct predictions by the system i.e. the value of true positive by our system is better and the value of incorrectly classified links i.e. false positive is less. Also the value of false negative i.e. the links that are not getting classified into appropriate class is 1 in many cases which increases the system performance. So the system is showing the better accuracy in case of F-measure value calculated from the precision and recall values obtained after submitting the keywords.

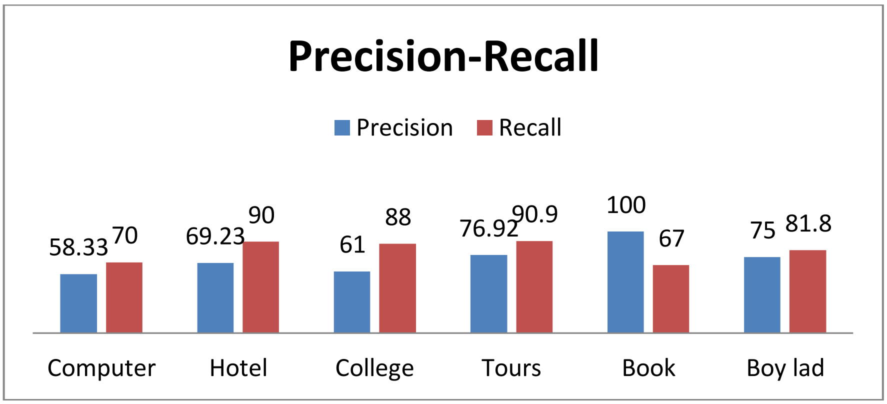

Fig.2 Precision and recall for various keywords

The precision and recall values obtained after submitting the various keywords to the search engine are calculated as shown in Figure 2. In all the cases we are getting the recall value greater than the precision value. When the keyword Book is submitted, all the links are classified correctly i.e. we are getting the false positive value as 0.So it is showing the 100% precision value for Book keyword. While in other cases the value of recall is greater than value of precision.

Precision is calculated as:

$$\text{Precision}=\frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}} \qquad (5)$$

Recall is calculated as:

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}} \quad (6)$$

From these precision and recall values the F-measure value is calculated as:

$$\text{F-measure}=\frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}} \quad (7)$$

Table1.F-measure values for different keywords

Keywords	F-measure value(%)
Computer	63.36
Hotel	78.26
College	72.05
Tours	83.34
Book	80.23
Boy lad	78.25

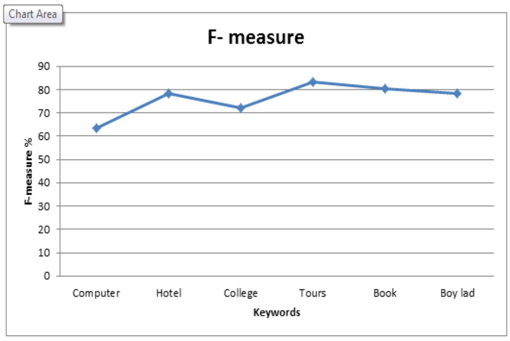

Fig. 3 F-measure values for various keywords

The accuracy is calculated as:

$$\text{Accuracy}=\frac{\text{Number of Correct prediction}}{\text{Total number of predictions}} \quad (8)$$

Table2. Accuracy for various keywords

Keywords	Accuracy (%)
Computer	47.00
Hotel	64.20
College	58.00
Tours	71.42
Book	67.00
Boy lad	64.28

The average of the accuracy calculated for each keyword submitted to the search engine is 61.98%.

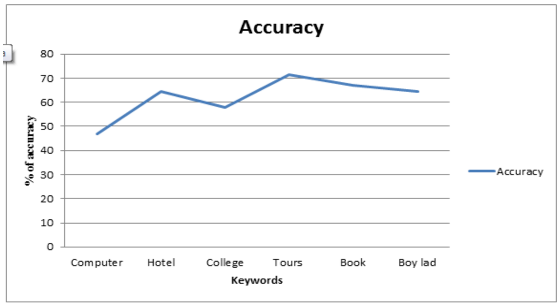

Fig. 4 Accuracy for keywords submitted

V. CONCLUSION

In the proposed system the various links are classifying into the predefined classes based on the highest probability of the word of each document. Among the all links extracted only few and relevant links are getting classified due to the use of Apriori algorithm. It is reducing the number of links to be classified that is reducing the complexity and time to scan for the all links. The web page classification system gives the F-measure value of 75.91% and gives the more efficiency and accuracy of classifier. We have found that the system works well even with large number of web pages used for classification. The value of true positive parameter for each keyword submitted is better while the value of false positive is less. So it is increasing the system performance whenever we submit any keyword to the system. The number of predefined class labels and number of words that are predefined for each class can be increased for getting the better accuracy in future.

References

1. Ajay S. Patil, B.V. Pawar" Automated Classification of Web Sites using Naive Bayesian Algorithm",Proceedings of International Multiconference of Engineering &Computer scientists 2012 Vol I,IMECS, Hong Kong,March 14-16,2012.
2. Igor Kotenko, Andrey Chechulin, Andrey Shorov,Dmitry Komashinsky "Analysis and Evaluation of Web Pages Classification Techniques for Inappropriate Content Blocking", ICDM 2014, LNAI 8557, pp. 39–54, Springer International Publishing Switzerland 2014.
3. Shiju Sathyadevan Athira U Sarath P R Anjana V "Improved Document Classification Through Enhanced Naive Bayes Algorithm", 2014 IEEE International Conference on Data Science and Engineering.
4. Keyur J. Patel1, Ketan J Sarvakar " Web Page Classification Using Data Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 7, July 2013.
5. Mohammed Al-Maolegi, Bassam Arkok "An Improved Apriori Algorithm For Association Rules", International Journal on Natural Language Computing (IJNLC)Vol. 3, No.1, February 2014.
6. Jiao Yabing" Research of an Improved Apriori Algorithm in Data Mining Association Rules" International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013.
7. Chowdhury Mofizur Rahman and Ferdous Ahmed Sohel and Parvez Naushad and S M Kamruzzaman, " Text Catagorisation using the Concept of Association Rule of Data Mining", CoRR 2010.
8. Abu Nowshed Chy ,Md. Hanif Seddiqui, Sowmitra Das "Bangla News Classification using Naive Bayes classifier", 16th International Conf. Computer and Information Technology, Khulna, Bangladesh, pp. 4799-3497, 8-10 March 2014.

AUTHOR(S) PROFILE

S. K. Dehankar has received her B.E. degree in Information Technology from Shri Sant Gajanan Maharaj College of Engineering, Shegaon Maharashtra, India in 2012.Currently pursuing M.Tech degree in Computer Science and Engineering from Government College of Engineering, Amravati, Maharashtra, India. Her research interest includes data mining. At present she is engaged with web document classification using data mining techniques.

K. P. Wagh has received his Diploma in Computer Engineering from Government Polytechnic Jalgaon. BE (CSE) from Government college of Engineering Aurangabad. ME (CSE) from Walchand College of Engineering Sangli. Presently he is working as Assistant Professor in Information Technology Department at Government College of Engineering, Amravati, Maharashtra.

Dr. P. N. Chatur has received his M.E degree in Electronics Engineering from Government College of Engineering Amravati, Maharashtra, India and Ph.D. degree from Amravari University. He has published twenty papers in international journals. His area of research includes Artificial Neural Network, Data Mining, Data Stream Mining and Cloud Computing. Currently, he is Head of Computer Science and Engineering & Electronics Engineering Department at Government College of Engineering Amravati, Maharashtra, India. At present he is engaged with large database mining analysis and stream mining.