International Journal of Management, Modern Sciences and Technologies
ISSN: 2305-199X || WWW.IJMMST.ORG
Open-Access || pp-10-14

Classification and Association of Web Pages Using Naïve Bayesian Algorithm

D V Nagarjana Devi, Dr. T. V. Rajanikanth, UlliTeja & Valluri Susmitha
Assistant Professor, RGUKT IIIT, Nuzvid.

Abstract: *World Wide Web (WWW), the information system is an effective standard for communication between computers. Search Engines are the sources to get any kind of information on the Internet. But they accomplish restricted ability in organizing the web pages. Nowadays classifying websites is very essential for easier access and associated information also very essential part to know more for internet users. There are two types of web page classification. First one is subject based web page classification and the other is genre based web page classification. Subject based classification can categorize the web pages into various categories and they store it in subject directories like Yahoo, Open Directory Projects, Directory Mozilla etc…based on the subject. Machine learningalgorithms can be used to automatically classify the websites. We focused on home pages of websites because they are the entry pointsand they provided all the links to the rest of the web pages. The information present in home page is an important source for classification like anchor tags, meta keywords, title and other information. Most of the web pages can have more information on home pages that can be very useful for classifying a web site into a particular category. In this paper we applied classification on content of the home pages of different categories using Naïve Bayesian algorithm and applying associationonclassified data using FPGrowth algorithm. Which will give effective associated websites based on given input.*

Keywords - *AssociationClassification, FPGrowthalgorithm,Machine learning, Naïve Bayesian algorithm.*

INTRODUCTION

Internet is the rich source to give access to the huge amount of data through search engines. This huge amount of data consists of vast amount of web pages also. Hence it is difficult to find out the target information for a user. Search engines can be of many types: several search engines are based on directory style such as Yahoo, some search engines are based on robot style such as altavista. Directory based search engines can store web pages in a database are classified in a hierarchical order. This enables classification easier. But it requires man power and takes much time and care.Moreover, the organization of these pages doesn't allow for easy search. So we need efficient and accurate methods to classify this huge amount of datainto categories But it is completely text based classification. There are numerous machine learning algorithms that have been used on text data for classification including Naïve Bayesian algorithm, Support Vector Machine, k-Nearest Neighbor algorithm and Neutral Networks.

In this paper we demonstrates Naïve Bayesian algorithm on content of the home pages for classification of web pages. Naïve Bayesian algorithm is one of the most successful and easy technique for text classification and gives fair results. Below mentioned details explains about paper organization.Section II analyses related work on the machine learning and classification of web pages. Section III provides clear information about the classification and association of the web pages. Section IV reviews Naïve Bayesian classification based on Bayes' theorem with some formulas. Section V discusses about the methodology of classification and association of web pages in an ordered manner. Section VI deliberates the results and outputs of our experiment. Section VII sums up the paper.

RELATED WORK

This section describes about the related works done on text classification of web pages. In past years classification was done by expertized persons. Few of the approaches proposed machine learning methods [3] for text categorization include Naïve Bayesian algorithm [1], support vector machine algorithm based on statistical learning theory [2], [4], k-nearest neighbor classification algorithm makes use of training documents, which have known classifications and discovers the closest neighbors of the new sample document amid all [11]. Other methods also include through reducing noise [12] it eliminates the noise in similarity measure, genre classification means classification based on the documents content [6], structures and procedures [7] define state-of-the-art practices, and trace the essential expectations overdue the usage of data from adjoining pages, integrating feature selection methods [8], feature intervals [9] the number of intervals each feature has to be discretized automatically.

In our proposed system, we demonstrate classification and association on content of the home pages of various categories using Naïve Bayesian algorithm and applied association on classified data using FP Growth algorithm.

CLASSIFICATION AND ASSOCIATION OF WEB PAGES

Web content classification is different when compared to text classification in some characterization. Additional challenges in classification of web pages introduced by the uncontrolled nature of web content as equated to customary text classification. The content of web comprises formatting data in the mode of HTML (Hypertext Markup Language) pages and it is semi structured.

 © Copyright **2013**. All Rights

Usually a web page contains hyperlinks which are pointing to other pages of different web sites. This kind of interconnection of web page gives characteristics that help greatly in classification of web pages. All HTML tags are taken out first in addition to punctuation marks from the web pages. In subsequent section we eliminate stop words which do not impart much in searching and also they are common to all documents. In most cases to decrease words to their basic stem we apply a stemming algorithm. Here we reduced the text by using text compactor software tool online. This is very easy to use and get accurate result from the given input. For the purpose of training the corresponding classifier machine learning algorithms are applied on those vectors. An unlabeled document against memorized data can be tested by algorithm classification mechanism. We described home pages of sports websites, organizational websites, banking websites and online shopping websites. For the whole website neatly designed home page is very much required. It gives the total view of entire website. More URLs connect to second level pages gives more information about characteristics of organization. The data included the title, meta description, labels and in meta keywords are most essential source of good features.

Site promoters provide a great amount of supply of similar keywords to rank high in search engine results. The extra data also can also be used effectively. In many cases home pages are created to match with a single screen. The characteristics described contribute to explain capacity of home page to recognize the organization nature. Association is an effective way for showing related information based on classified data. It is not a difficult task and it can encourage the user to know more about the organization or category to proceed further.

NAÏVE BAYESIAN CLASSIFICATION

Naïve Bayesian (NB) classification is built on Bayes' theorem with individuality assumptions among predictors. NB model or classifier is very easy to build. There is no complex iterative parameter estimation in this approach which makes it mainly useful for very large data sets. This is unique and successful known algorithm for classification of text documents. A term-class grouping that does not ensue in the training data makes the complete outcome zero. In order to resolve this problem we can use add-one smoothing or Laplace smoothing for better performance.

NB approach is appropriate when the dimensionality of the input is high. It requires slight amount of training data to evaluate the parameters.

For instance:

D: Set of tuples
- Each Tuple is an 'n' dimensional attribute vector
- X: $(x_1, x_2, x_3 \ldots x_n)$

Let there be 'm' Classes: $C_1, C_2, C_3 \ldots C_m$

Naïve Bayes classifier predicts X belongs to Class Ci iff
- $P(C_i/X) > P(C_j/X)$ for $1 <= j <= m, j <> i$

Maximum Posteriori Hypothesis
- $P(C_i/X) = P(X/C_i)\, P(C_i) / P(X)$
- Maximize $P(X/C_i)\, P(C_i)$ as $P(X)$ is constant

With many attributes, it is computationally expensive to evaluate $P(X/Ci)$.

Naïve Assumption of class conditional independence

$$P(X / C_i) = \prod_{k=1} n\, P(x_k / C_i)$$

$$P(X / C_i) = P(x_1 / C_i) * P(x_2 / C_i) * \ldots * P(x_n / C_i)$$

$P(C_i)$, the probability of c is calculated as:

$$P(C_i) = N_c / N$$

Where

N_c = Number of training documents in class c

N = Number of training documents

$P(c \mid x)$ = Subsequent probability of cas it reproduces our confidence that c grips afterwards we have understood x.

METHODOLOGY

This section briefly discusses about the overall concept of this experiment. First of all we require gathering of allhome pages of web pages which are pre classified into different categories. These pre classified web pages can be cleaned by removing HTMLtags and scripts, stop words present in each web page. Then the data set can give it to NB classifier for training and testing the classifier. The whole thing can be explained below.

The architecture of the proposed system

The general design of our project is shown in Fig.1 below. The data is collected directly from various search engines like Google, Yahoo, Bing etc,.in a html format then pre-process the html file into text file to move further process. After that we perform Naïve Bayesian classification on text data for classification and applying association on classified data. The following steps are clearly defined the proposed system.

 © Copyright 2013. All Rights

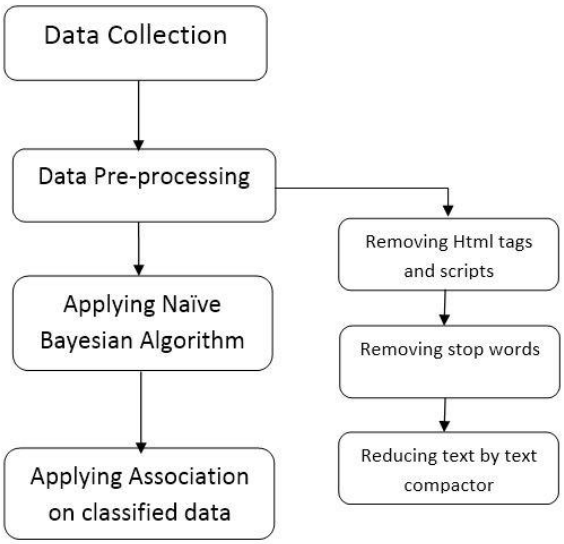

Fig.1. The design of the proposed system

Algorithm:
Step1: Collection of data set
Step2: Data Pre-processing
2.1 Removing HTML tags and scripts
2.2 Removing stop words
2.3 Reducing text using Text Compactor
Step3: Applying Naïve Bayesian classification on text data
Step4: Applying association algorithm on classified data

1) Data Collection
Data collection is the process ofcollectingweb pages in our project. There are several tools to collect the data from search engines. But in this project we have directly collected the data without using any tool in a html format. Our text file contains different categories of web pages regarding Universities, Banking systems, Sports category, Online shopping information and Payment systems etc.. We focused on each category of different web pages which are mentioned above and related issues are considered.

2) Data Pre-processing
2.1) Removing html tags and scripts
In addition to the actual information of web pages, contain other meta-data information. Here weare having the data in a html format. HTMLAsText is software which converts HTMLdocuments to simple text files.It willremove all HTML tags and formatting the text according to your preferences.It will automatically remove alltags and scripts from the HTML document. The remaining text can be formatted based on text that you select. The extracted text can stored onto a file for further process.

2.2) Removing stop words
The input text can be read line by line by using FileInputStream class in Java. Java code can do the pre-process as removal of Punctuation marks and removal of stop words like home, search, login, view, web master, subscribe and feedback etc. Because we have to classify the text based the content present in home page. We don't need this kind of words for classification. So here we considered this kind of words as stop words. It can give you output as the new text file after java code had been executed.

2.3) Reducing text by using Text Compactor
Text Compactor is a software tool which will reduce the text based on given input online.We need to place the text on the page, the web automatically calculates the frequency of each wordin the passage. We need to place the text on the page, the web repeatedly evaluates the frequency of each term in the passage. Human readers may not agree with this automated methodology to text summarization. It works well on expository text such as textbooks and reference material. When a passage has only a limited sentences then the results can be tilted.

3) Applying Naïve Bayesian Classification
Data collection is a gathering of web pages in our project. There are several tools to collect the data from search engines. But in this project we are collected the data directly from internet without using any tool in a html format. Our text file contains different categories of web pages regarding Universities, Banking systems, Sports category, Online shopping information and Payment systems etc.. We focused on each category of different web pages which are mentioned above and related issues are considered.

 © Copyright **2013**. All Rights

To test the classifier we need training data set and test data set. To train the data some set of training examples are used for classifier. For testing classifiers are applied to classify web pages. 60 percent of input data can be used for training the classifier and remaining 40 percent input data can be used for testing the classifier. Training and test data can in a text format in a text file. That can be given as input to our Naïve Bayesian classifier. Our Naïve Bayesian algorithm is written in pythoncode. Python is simple to write and works well for data mining algorithms. Based on training and test data our python code can give output in a text file contains classified web pages text. Classified textcan be used for association algorithms to generate associated web sites.

4) Applying Association algorithm on classified data

Association can be applied by association rule mining algorithms. There are so many algorithms to apply association on text. Association rules can be used to find out the association among large set of data. Coming to detailed information minimum support threshold in addition to minimum confidence threshold are robust and they are used to satisfying the association rules. Set of items we call it as items sets. Set of items which contain k items is k-item set. Number of transactions that comprises the item set referred as frequency occurrence. Also known as support count or simply count. An Item set should satisfies minimum support threshold then it is a frequent item set.

In our project we used FPGrowth algorithm for association of web pages. FPGrowth algorithm is one of the effective method to generate association rules. To generate rules it will follow two steps. First it will find all frequent item sets after that it will generate association rules from frequent item sets. FPGrowth algorithm is written in java code. It will take text file as input and gives text file as output. Input data contains classified web pages data from the Naïve Bayesian algorithm. Nearly 4500 classified web pages can be given as input in a text file to FPGrowth algorithm. It will give output as frequent item set with support count or count. The FPGrowth algorithm can gives output based input text file.

EXPERIMENTAL RESULTS

We examined our implementation of this approach over a sample space about 4500 pages. From various domains these papers are collected. In Table 1 the results and various parameters used are written.

We got some results from the implementation we have done and we found that some results were not sync with the original category for which paper belongs to. Consider an example shown in figure 2. We can observe that the text shown in the figure has to fall in information image but not completely. The reason can be explained by the fact that it is actually print version of original one and there is no a single link in the page.

Detail Information	**Results**
No. of pages on which we have tested our implementation	~4500
Pages categorized	~3950
Pages categorized correctly	~3520
% categorized correctly	89.11%

Table 1

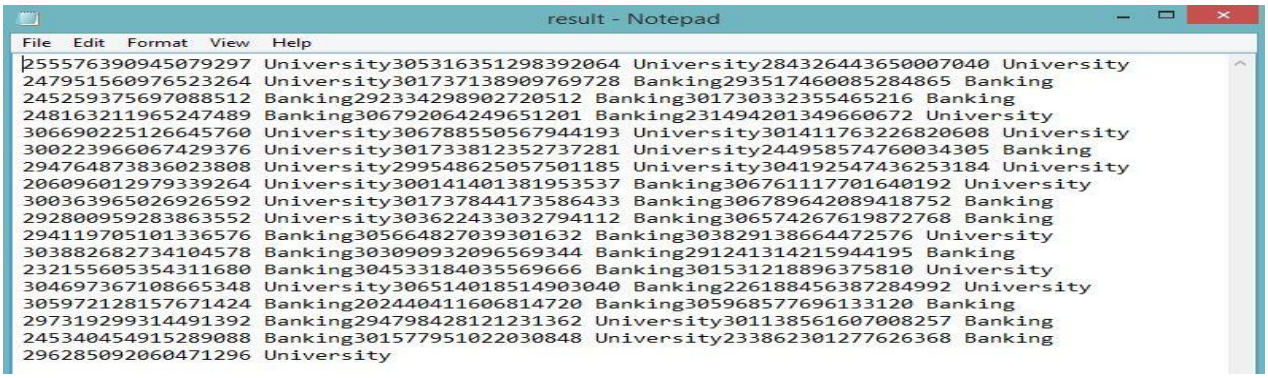

Fig.2. Categorized web sites

For association of web pages we used FPGrowth algorithm as we mentioned earlier. For FPGrowth Algorithm we have given nearly 3500 domain names as input file which we pre classified before using Naïve Bayesian algorithm. Domain names or classified web pages can be store in a text file which is shown in below figure 3.

 © Copyright **2013**. All Rights

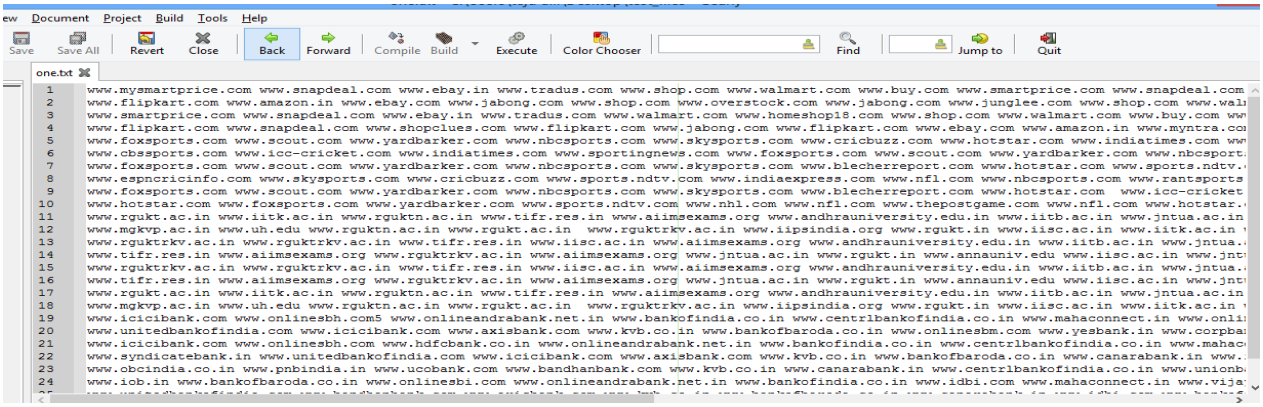

Fig.3. Input data given for FPGrowth Algorithm

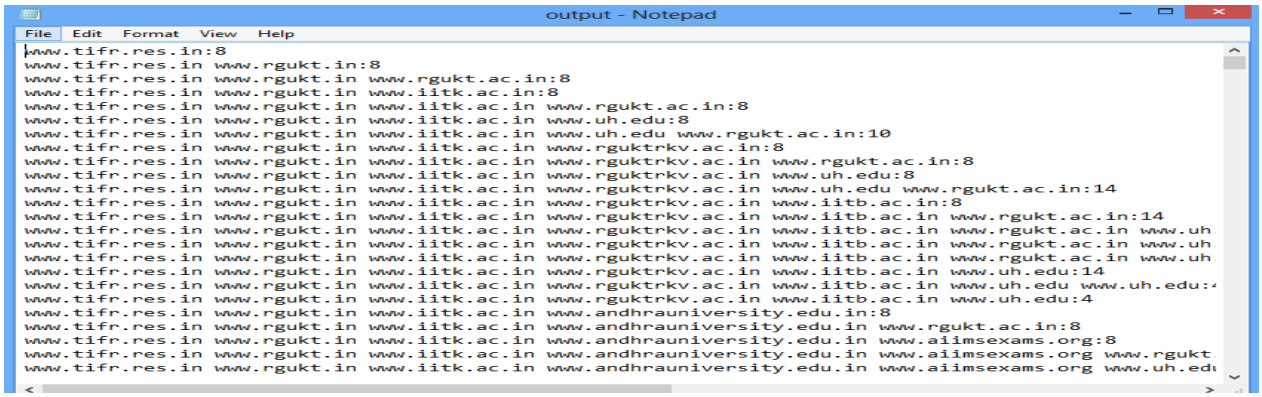

Fig.4. Output data obtained from FPGrowth Algorithm

Based on input given FPGrowth can generate association rules and frequently identified items with support count or simply say count. It will check input data gives the result in a text file which is shown in below figure 4.

CONCLUSION AND FUTURE WORK

Classification and Association of web pages is still interesting problem in general fields. In this paper we classified the web pages based on NB classifier for effective categorization. It classifies the web pages into numerous categories. The classification of web pages for the ten categories recorded above are using NB approach conceded 89.11% accuracy. We perceived that the classifier's classification accuracy is directly proportional to the number of documents which are trained by NB classifier. The results are more boosting. For association of web pages FPGrowth algorithm gives good results. This concept can be used by search engines for effective classification of web pages and its association. The distinct and non hierarchical categories are studied in this experiment. This approach can be used for classify the web pages into more explicit categories with its associated data which can be useful for the users more clear.

REFERENCES

- *Ajay S. Patil, BV Pawar, Automated Classification of Websites using Naïve Bayesian Algorithm, 2012.*
- *Web PageClassification by Ben Choi and Zhongmei Yao.*
- *Makoto Tsukada, Takashi Washio, Hiroshi Motoda, Automatic Web-Page Classification by using Machine Learning Methods.*
- *W.A.AWAD, Machine Learning Algorithms in Web Page Classification, International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 5, October 2012.*
- *Arul Prakash Asirvadham, Kranthi Kumar. Ravi, Web Page Categorization based on Document Structure.*
- *Sven Meyer zuEissen and Benno Stein, Genre Classification of Web Pages.*
- *XIAOGUANG QI and BRIAN D.DAVISON, Web Page Classification: Features and Algorithms, Feb 2009.*
- *J. Alamelu Mangai, Dipti D. Kothari and V. Santhosh Kumar, A Novel Approach for Automatic Web Page Classification using Feature Intervals, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 2, September 2012.*
- *A. KousarNikhath,K.Subrahmanyam, R.Vasavi, Building K-Nearest Neighbor classifier for text categorization, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (1) , 2016.*
- *LI Xiaoli and SHI Zhongzhi, Innovating Web Page Classification through Reducing Noise, 2002.*

 © Copyright **2013**. All Rights