WEB PAGE CLASSIFICATION USING TERM FREQUENCY

Ms. Sonal Vaghela[1], Mr. M.B.Chaudhary[2], Mr. Devendra Chauhan[3]
[1, 2]Computer Engineering Department
Government Engineering College, Gandhinagar, Gujarat, India.
[3]Computer Engineering Department,
Balasinor College of polytechnique, Gujarat, India.

Abstract: World Wide Web contains a lot of web pages. Web pages contain the various types of information. Based on web page information we have to classify the web page. Web page classification is area of web mining. Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. Web Page Classification can done based on different feature selection methods like term occurrence number, term frequency, Document frequency etc. This research paper is focuses on the Web Page Classification using the term frequency.
Keywords: Web Page classification, Features, Classifiers, Naïve Bayes, Support Vector Machine, kNN.

I. INTRODUCTION

The fast developments on the computer and networking technologies have increased the popularity of the Web which has caused the inclusion of more and more information on the Web [10]. Web contents, including online documents, e-books, journal articles, technical reports and digital libraries, have been rapidly exploring all time. It is much helpful to categorize web contents for efficiently contents browsing, managing, even spam filtering [5][16]. Web page classification, also known as web page categorization, is the process of classifying the web pages into the predefined categories. Classification is one of the traditional data mining tasks. Classification is often posed as a supervised learning problem in which a set of labeled data is used to train a classifier which can be applied to label future examples.

According to Qi and Davison [2], the general problem of web page Classification can be divided into multiple sub-problems: Subject Classification concerns about the subject or topic of web page. Functional Classification cares about the role web page plays. Sentimate Classification focuses on opinion presented in the web page. Binary Classification categorizes each instance into one of the two categories. Multi-class Classification deals with more than two classes. Multiclass Classification can be further divided into single-label a multi-label classification. Flat Classification in that categories are considered parallel, i.e., one category does not supersede another. Hierarchical Classification the categories are organized in a hierarchical tree-like structure, in which each category may have a number of subcategories [6]. Web page Classification is a process where one page is appended to one or more directories which is predefined in advance [3]. Automatic Web page classification is a supervised learning problem in which a set of labeled Web documents is used for training a classifier, and then the classifier is employed to assign one or more predefined category labels to future Web pages [2]. Most of the applied web page classification techniques are inherited from automatic text classification: a supervised learning task, defined as assigning pre-defined category labels to new documents, based on the likelihood suggested by a training set of labeled documents. Therefore, an increasing number of learning approaches have been applied to classify web pages [11] [16]. The rest of the paper is organized as follows. Web page classification applications are given in section II. Web page classification methods are presented in section III. Section IV covers the related work. Section V shows the proposed classification system. Section VI covers the Experimental results and Section VII Conclusions. Finally Section VIII shows future Work.

II. WEB PAGE CLASSIFICATION APPLICATIONS

Applications of Web Page Classification are [6]:

- Constructing, maintaining or expanding web directories (web hierarchies)
- Improving quality of search results
- Helping question answering systems
- Building efficient focused crawlers or vertical search engines
- Web content filtering
- Assisted web browsing
- Knowledge base construction

III. WEB PAGE CLASSIFICATION METHODS

Web Page can be classified into the following broad categories [6] [8]:

- Manual classification
- Clustering approaches
- META tags based categorization
- Text content based categorization
- Link and content analysis

In addition to these, document frequency, term occurrence number based approach has also been used for classifying web pages [15]. Among these methods the work presented in this paper is based on the combination of structure and text content approaches using Term frequency, hence it is named web page classification Using Term Frequency.

IV. RELATED WORK

[10] Presented genetic algorithm based automatic web page classification approach which uses both the HTML tags and terms belong to each tag as classification features. Feature extraction part takes positive examples in the training dataset and determines features that are used in the coding process of the GA. The system classifies Web pages by simply computing similarity between the learned classifier and the new Web pages. They used three datasets, conference, and student and course, each contains positive and negative examples. They classified whether a particular example is positive or negative in the dataset based on learned classifier. A research [17] presents web page classification technique based on the data extracted from the HTML code. In that the data processing selects from the raw data base a data set that focuses on a subset of attributes or variables on which knowledge discovery is to be performed. It uses HTML code to represent the processed data by means of an Object Attribute Table (OAT). The OAT contains the columns like Page's text length (TL), External links (EL), Internal links (IL), Image (Im), External Images (EI), Internal Images (II),Multimedia Objects (MO), Word Flash (WF), Word Video(WV), Word Image (WI), Word Blog (WB), Word News (WN). Each row of the OAT describes the characteristics of a web page using these defined attributes. In the following step, an expert assigns classes to each of the rows according with the four categories, Blog, News, Video and Image. In the data mining phase, decision trees converts the data contained in the OAT into useful patterns. In the evaluation phase the consistency of pattern is proven by means of a testing set.

V. CLASSIFICATION SYSTEM

The methodology presented here uses the Document structure i.e. HTML tags of web page together with the contents within it which is the traditional text categorization approach.

A. Feature Extraction

In the feature extraction stage, candidate features (i.e., the original feature set) are generated from the training set. Among the various HTML tags, <title>, <h1>, <h2>, <h3>,<a>, , <b>, , <i>, <p>, and <li> tags which denote title, header at level 1, header at level 2, header at level 3, anchor, strong, bold, emphasize, italic, paragraph, and list item, respectively, contains most of the domain specific important terms. So it will be beneficial to consider these tags to generate features that are used in both classifier learning and classification processes. To generate features, all the terms from each of the above mentioned tags are taken then; stop word are removed and Porter's stemming [18] algorithm are applied. Each stemmed term and its corresponding tag form a feature. As an example the word "web" in <title> tag, "web" in <b> tag and "web" in <li> tag are considered as different features. The feature extraction algorithm is shown in Fig 1.

Algorithm 1 : FEATURE EXTRACTION

Input: Collection of Web pages, Stopword list.
Output: Features list

```
for each Web page p in collection do
  for each word w in p do
    if w is not stopword then
      if w belongs to <title> tag then
        title = title U stem(w)
      else if w belongs to <h1> or <h2> or <h3> tag then
        header = header U stem(w)
      else if w belongs to <a> tag then
        anchor = anchor U stem(w)
      else if w belongs to <em> or <strong> or <b> or <i>tag then
        bold = bold U stem(w)
      else if w belongs to <li> tag then
        list_item = list_item U stem(w)
      else if w belongs to <p> tag then
        paragraph = paragraph U stem(w)
      end if
    end if
  end for
end for
```

Fig. 1: Feature Extraction Algorithm

B. Feature Selection

Feature selection/Feature extraction is an important in pattern recognition or pattern classification, data mining, web mining and machine learning. It also helps to remove noise features in web pages so as to improve search efficiency. The basic idea of feature selection algorithms is searching through all possible combinations of features in the data to find which subset of features works best for prediction and hence in searching. The selection reduces the number of features, keeping the most meaningful ones, and removing the irrelevant or redundant features. Various methods for feature selection are [19]: (1) Term occurrence number (2) Term frequency (TF) (3) Document frequency (DF) (4) Inverse Document Frequency (IDF).

C. Classifier

There are various Machine learning algorithms available for classification:
Decision Tree, Naive Bayes, Neural Network, Support Vector Machine, and K- nearest neighbour etc. This work used Naïve Bayes and Support Vector Machine and K-nearest neighbour for web page classification.

1) Naïve Bayes:
Bayesian learning is a probability-driven algorithm based on Bayes probability theorem as the follow [10]:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

Where X is considered "evidence", which is described by measurements made on a set of n attributes. and H is some hypothesis, i.e. data tuple X belongs to a specified class C.

P (H|X) is the posterior probability of H Conditioned on X. That is tuple X belongs to class C, given that we know the attribute description of X that is we want to determine for classification problems.

P(H) is the prior probability of H.
P(X|H) is the posterior probability of X conditioned on H.
P(X) is the prior probability of X.

The NB works as follows: Each data sample is represented by an n-dimensional feature vector, X=(x1,x2,…,xn), representing n measurements made on the sample from n attribute, respectively, A1,A2,…,An. Suppose that there are m classes, C1,C2,…,Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditional on X. That is, the naive Bayesian classifier assigns an unknown
sample X to the class Ci if and only if:

$P(Ci|X) > P(Cj|X)$, where $1 \le j \le m$, and $j \ne i$.
The class for which P (Ci|X) is called as the maximum posteriori hypothesis.

By the Bayesian theorem:

$$P(Ci|X) = \frac{P(X|Ci)\,P(Ci)}{P(X)}$$

As P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are unknown, then it is commonly assumed that the classes are equally likely, that is: P (C1) = P (C2) = ….. = P (Cm)
Note that the Class prior probabilities may be estimated by

$$P(Ci) = \frac{Si}{S}$$

Where, Si is the number of training samples of class Ci and S is the total number of training samples.

In order to reduce computation in evaluating, the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample that is there are no dependence relationships among the attributes. Thus, In order to classify unknown samples X, P (X|Ci) P (Ci) is evaluated for each class Ci. Sample X is then assigned to the class Ci based on the higher posterior probability.

2) Support Vector Machine:
Support Vector Machines are among the most robust and successful classification algorithms. It is a new classification method for both linear and nonlinear Data. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [10].

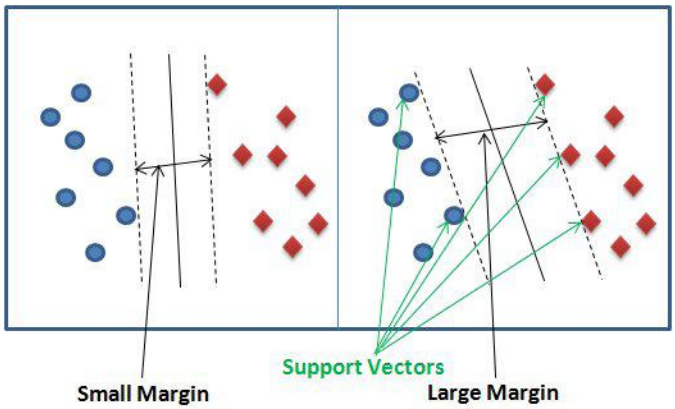

Fig. 2: Support Vector Machine

SVM uses a nonlinear mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyperplane (i.e. decision boundary). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. There are many hyperplanes that might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. So the best hyperplane is the one that the distance from it to the nearest data point on each side is maximized. If such a exists, it is known as the maximum-margin
hyperplane as shown in Fig 2. SVM have several advantages. Because the margin maximization and the regularization term, SVM are known to have high accuracy, good generalization properties, to be insensitive to overtraining and to the curse-of-dimensionality. These advantages are gained at the expense of a low speed of execution [10].

3) K- Nearest Neighbour
For the classification, there several algorithms used, one of which is K-Nearest neighbors. Nearest neighbor search is one of the most popular learning and classification techniques, which had been proved to be a simple and powerful recognition algorithm. It finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular

class in this neighborhood [20]. Nearest-neighbor classifiers are based on learning by analogy, in that it comparing a given test tuple with training tuples. The training tuple are related to n attributes. Each tuples denotes a point in an n-dimensional space. In this way, all of the training tuples are denoted in an n-dimensional pattern space. When given an unknown tuple, k-nearest-neighbor classifiers find the pattern space for the k-training tuples that are nearer to the unknown tuple. These k-training tuples are the k "nearest neighbors" of the unknown tuple [3].

D) Dataset

The dataset contains the web pages related to each of the categories used for classification. As the categories are conference page, student home page, department and course, dataset contains the web pages for Computer Science related conference homepages, graduate students' homepages, department pages and Computer Science course homepages, respectively. There are total 201 web pages are used for this classification work. Among them there are 40 web pages of conference, 61 of student home page, 39 of department pages and 61 are of course pages. Computer Science related conference homepages that were obtained from "Computers: Computer Science: Conferences" category of the Open Directory Project web site (http://www.dmoz.org).The Course, Department and the Student web pages taken from well-known and freeware datasets that were obtained from the WebKB project Website (http://www.cs.cmu.edu/ webkb).

VI. EXPERIMENTAL RESULTS

Data mining tool WEKA is used to perform the classification. The classification is done with 201 examples and with different values of term frequency. The classifiers' performances has been analyzed and compared by the measures Precision, Recall and F-measure, which are obtained from the confusion matrix.

CONFUSION METRIX:

TABLE I
CONFUSION MATRIX

	Category 1	**Category 2**
Classified as 1	True Positive	False Positive
Classified as 2	False Negative	True Negative

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\text{F – measure} = \frac{2*\text{Precision}*\text{Recall}}{(\text{Precision + Recall})}$$

TABLE II
PRECISION, RECALL AND F-MEASURE

	NB			**SVM**			**kNN**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Conference	0.97	0.8	0.877	0.8099	0.95	0.8744	0.3700	1.000	0.541
Course	0.894	0.967	0.929	0.9666	0.918	0.941	1.0000	0.4922	0.659
Department	0.974	0.974	0.974	0.974	0.949	0.961	0.9666	0.7188	0.824
Student	0.937	0.967	0.952	0.9655	0.902	0.932	0.9666	0.7188	0.824
AVG.	0.929	0.926	0.925	0.927	0.922	0.921	1.0000	0.5557	0.716

Different Term frequency values were experimented. The Term frequency vs. Classification Accuracy and Term frequency vs. Time graphs are shown in Fig 3. and Fig 4., respectively. Table II shows Precision, Recall and F-measure values for term frequency value 4.

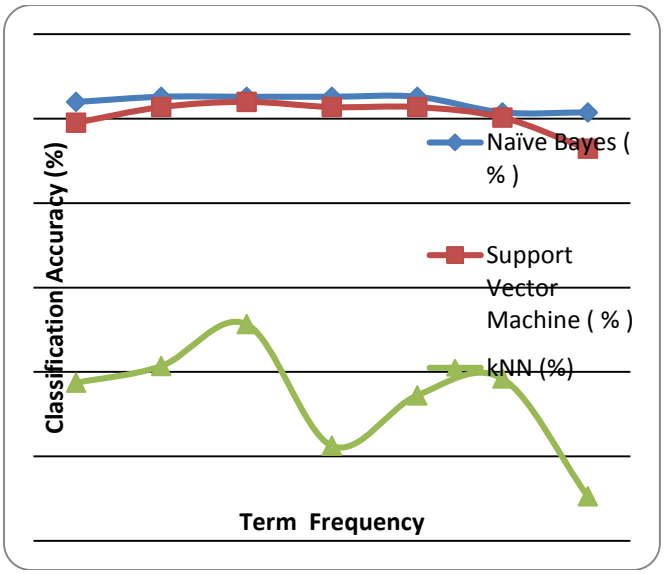

Fig. 3: Term Frequency vs. Classification Accuracy

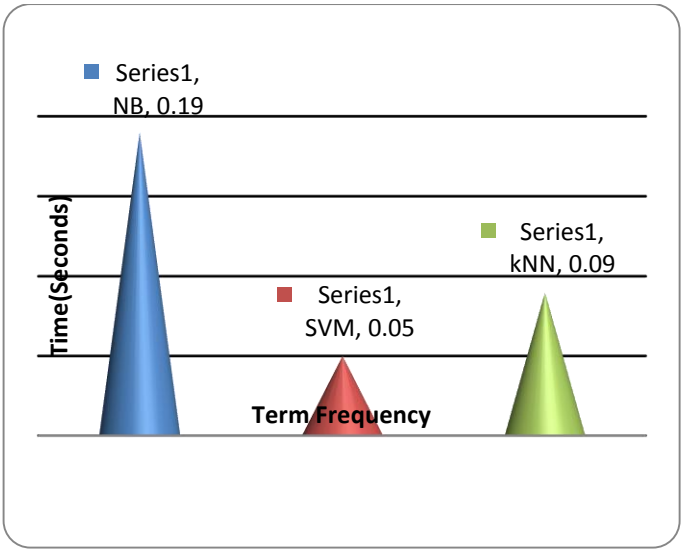

Fig. 4: Term Frequency vs. Time

VII. CONCLUSIONS

Web page classification with html tag and term within each tag - combination as features classifies web pages more accurately. Naive Bayes classifier performs the classification with 92.54 % accuracy, and SVM performs classification with 91.54% accuracy, while kNN performs the classification with 65.67%. Thus it is concluded that Naive Bayes is good for web page classification with combination of html tag and term as feature. This work classifies web pages of four categories conference, course, department and student. Thus it can be used for educational universities' catalogs. This catalog provides users categorized view of information and is more effective for users to find desired information. So Users of universities can easily find information about conference, course, department and students. Building such catalog manually require lot of human effort.

VIII. FUTURE WORK

In this dissertation work, the web pages are classified into four categories, conference, department, student and course. Thus it can be used for developing, maintaining Educational Universities' catalogs, which can be expanded by including more categories such as, projects, Event, staff etc. As this work is related to classification of educational web pages, analyzing them specifies the fact that most web pages contain copyrights and various images as noise. Noise is somewhat removed but not completely which can be further improved. A small set of tags is considered for feature generation. META tags, images, scripts etc. are not considered which can be included. The Classification algorithm used for this dissertation work is Naïve Bayes, Support Vector Machine and k- Nearest Neighbour. So, other Classification algorithms can also be implemented. The Feature selection method which has been used for this dissertation work is Term frequency. Other methods such as Term Occurrence number, Document frequency (DF), IDF can be used.

IX. ACKNOWLEDGEMENT

I would like to acknowledge the My Sir, Prof. M.B. Chaudhari for his kindness and support to me for doing my research work and to my husband and my family for allowing me to snatch the time of my life which they want to spend with me.

REFERENCES

[1] J.Krutil, M.Kudelka and V. Snasel,"Web Page Classification based on Schema.org Collection", In: Fourth International Conference on Computational Aspects of Social Networks(CASoN),2012

[2] Sarac, E.;Ozel, S.A.," Web Page Classification Using Firefly Optimization", In: Innovations in Intelligent Systems and Applications(INISTA),2013 IEEE International Symposium, pages 1-5

[3] Wang Zhixing and Chen Shaohong,"Web Page Classification based on Semi-supervised Naive Bayesian EM Algorithm",In: IEEE International Conference on Communication Software and Networks(ICCSN),2011 pages,242-245

[4] Yusuf , L. M. and Othman, M.S., Salim, J., "Web Classification using Extraction and Machine Learning Techniques", In: Information Technology(ITSim)2010 International Symposium in (volume:2),

[5] Tian Xia,Yanmei Chai,Tong Wang,"Improving SVM on Web Content Classification by Document Formulation", In: 7th International Conference on Computer Science & Education(ICCSE 2012)

[6] D.Navadiay,M.Parikh,R.Patel,"Constructure Based Web Page Classification", International Journal of Computer Science and Management Research, Vol 2,Issue 6, June 2013

[7] M. IndraDevi, Dr. R. Rajaram, K. Selvakuberan ,"Automatic Web Page Classification by Combining Feature Selection Techniques and Lazy Learners", In: International Conference on Computational and Multimedia Applications (ICCIMA),2007

[8] A. Asirvatham, K. Ravi," Web Page Categorization based on Document Structure", www.iiit.ac. In /~arul/-paper.pdf.

[9] ZHI-MING XU,XIN-BO GAO,MENG LEI,"WEB SITE CLASSIFICATION BASED ON KEY RESOURCES", In: Proceedings of the 8th International Conference on Machine Learning and Cybernetics,2009.

[10] Selma Ayse Ozel,"A Web page Classification System Based on a genetic algorithm using tagged-terms as feature", In: Journal On Expert System Applications 38(2011)3407-3415.

[11] V.Fernandez,R.Unanue,S.Herranz,A.Rubio,"NaıveBayesWebPageClassificationwithHTMLMark-UpEnrichment",In Proceedings of the International Multi-Conference on computing in the Global Information Technology, 2006.

[12] S. Meher, S. Pal, S. dutta, "Granular Computing Models in the Classification of Web Content Data", In International Conferences on Web Intelligence & Intelligent Agent Technology, 2012.

[13] S. Balan, " A study of various Techniques of Web Content Mining Research Issues and Tools", In International Journal of Innovative Research and Studies,Vol-2,Issue-5,May-2013.pp508-517
[14] Cooley R., Mobasher B, Srivastava J," Web mining: information and pattern discovery on the World Wide Web",. In Proceedings of Ninth IEEE International Conference.Nov-1997.Pp558-567.
[15] D.Navadiya, R.Patel,"Web Content Mining Techniques-A Comprehensive Survey", International Journal of Engineering Research & Technology (IJERT), Vol.1 Issue 10, December-2012, Pp1-6.
[16] S.Vaghela, P.Patel," Web Page Classification Techniques-A Comprehensive Survey", International Journal of Engineering Development & Research (IJEDR),Vol.1 Issue 3,December-2013,Pp228-230
[17] Gabriel Fiol-Roig, Margaret Miro-Julia, Eduardo Herraiz, "Data Mining Techniques for Web Page Classification", 2011.
[18] Porter, An algorithm for suffix stripping. Program, 14(3), 1980, 130-137.
[19] Chih-Ming Chen, Hahn-Ming Lee, Yu-Jung Chang,"Two novel feature selection approaches for Web page classification", Expert Systems with Application 36, 2009, 260-272.
[20] Putu Wira Buana, Sesaltina Jannet D.R.M, I Ketut Gede Darma Putra," Combination of K-Nearest Neighbor and K-Means based on Term Re-weighting for Classify indonesian News", International Journal of Computer Applications (0975–8887)Vol.50 –No.11,July 2012

