Web Page Classification Using Feature Selection Techniques

[1]Smt. P. G. Modh [2]Mr. M. B. Chaudhari
[1]ME Scholar, [2]Professor
Department of Computer Engineering, Government Engineering College, Gandhinagar

***Abstract* - Websites provides a lot of information to the Users. Websites is a collection of Web Pages which contains a bunch of information. In these information to find or retrieve particular page or information is difficult task. So to make this task easy there are different web page classification methods. Using this method we can identify web pages. 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 WWW. Web page Classification retrieves WebPages based on different features/parameters of web pages. Web Page Classification can be done in many ways such as using the content of the web page, using the structure of web page, using clustering methods etc. This paper is focuses on the Web Page Classification using the inverse document frequency.**

***Index Terms* - Web Page classification, Classifiers, Features, Naïve Bayes, SVM**

I. INTRODUCTION

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 [1], 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 an 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.

II. WEB PAGE CLASSIFICATION APPLICATIONS

Applications of Web Page Classification are [1]:

- 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]:

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

IV. RELATED WORK

[3] 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 [2] 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.

A paper [7] used the HTML form elements and their attributes to classify the Web pages with HTML forms into External search forms, No search, site search, Internal database search using the random forest classifier. In order to eliminate sparse matrix problem and to avoid preprocessing of strings, numeric data such as the numbers of input element text type, checkbox type, radio type, select element and textarea element in a form were gathered. In addition, all strings from label elements and name attribute and value attribute of input type text, checkbox and radio, element select and element textarea were scanned for word "search" and its synonyms.

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.

1. *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; stopword are removed and Porter's stemming [4] 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.

Algorithm 1: FEATURE EXTRACTION ALGORITHM

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 or 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

2. *Feature Selection*

Feature selection/Feature extraction is an important pre-processing step 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 [8]: (1) Term occurrence number (2) Term frequency (TF) (3) Document frequency (DF) (4) Inverse Document Frequency (IDF)

Inverse Document Frequency (IDF)

- It is the logarithmically scaled fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient**.**

$$IDF(t) = \log(\text{Total number of documents}) / (\text{Number of documents with term t in it})$$

3. *Classifier*

There are various Machine learning algorithms available for classification:

Decision Tree, Naive Bayes, Neural Network, Support Vector Machine, K- nearest neighbour etc. This work used Naïve Bayes and Support Vector Machine for web page classification.

A) Naïve Bayes :

Bayesian learning is a probability-driven algorithm based on Bayes probability theorem as the follow[9]:

$$P(H|X) = P(X|H)\, P(H) / P(X)$$

Where X is considered "evidence", which is described by measurements made on a set of n attributes. and H be 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:

$$P(Ci|X) > P(Cj|X) \text{ , where } 1 \leq j \leq m, \text{ and } j \neq i.$$

The class for which P(Ci|X) is called as the maximum posteriori hypothesis.

By the Bayesian theorem :

$$P(Ci|X) = P(X|Ci)\, P(Ci) / P(X)$$

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

$$P(Ci) = Si / S$$

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

B) 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.

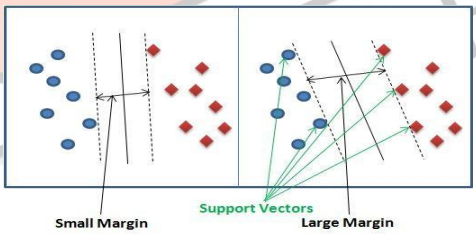

Fig 1. Support Vector Machine

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 hyperplane exists, it is known as the maximum-margin hyperplane as shown in Fig 1.

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.

4. *Dataset*

The dataset contains the web pages which we refer as a documents related to each of the categories used for classification. As the categories are course page, faculty page, project page, student page, dataset contains the web pages for Computer Science related course homepages, faculty homepages, project pages and Computer Science related student homepages, respectively. There are total 1903 web pages are used for this classification work. Among them there are 564 web pages of course, 522 of Faculty home page, 269 of Project pages and 548 are of Student pages. The Course, Faculty, Project and the Student web pages taken from well known and freeware datasets that were obtained from the WebKB project Web site (http:// www.cs.cmu.edu/ webkb).

VI. EXPERIMENTAL RESULTS

Data mining tool WEKA is used to perform the classification. The classification is done with 1903 examples and with different values of Inverse Document Frequency. The classifiers' performances has been analysed 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

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

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

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

Different Inverse Document Frequency values were experimented. Table II shows the classification accuracy of NB and SVM algorithms with different values of Inverse Document Frequency and corresponding number of attributes. The table shows that the maximum result obtained for both algorithms, is with Inverse Document Frequency 4. The Inverse Document Frequency vs. Classification Accuracy and Inverse Document Frequency vs. F-measure graphs are shown in Fig 2. and Fig 3., respectively. Table III shows Precision, Recall and F-measure values for Inverse Document Frequency value 4.

TABLE II CLASSIFICATION ACCURACY

Inverse Document Frequency	No. of Features	Naïve Bayes	Support Vector Machine
1	2838	88.4752	93.4397
2	1945	89.0071	93.7943
3	1644	89.3617	93.7943
4	**1380**	**89.539**	**93.7943**
5	1250	88.8298	94.1489
10	1086	88.8298	93.4397
15	952	89.0071	93.2624
17	867	88.6525	93.4397
20	785	88.4752	93.4397
25	655	89.0071	93.2624
30	566	89.1844	92.3759

TABLE III PRECISION, RECALL AND F-MEASURE

	NB			SVM		
	P	R	F	F	R	P
Course	0.87	0.885	0.87	0.923	0.834	0.926
Faculty	0.882	0.858	0.861	0.949	0.956	0.958
Project	0.804	0.881	0.839	0.834	0.881	0.804
Student	0.924	0.903	0.913	0.969	0.969	0.969
AVG	0.87	0.88175	0.87075	0.91875	0.91	0.91425

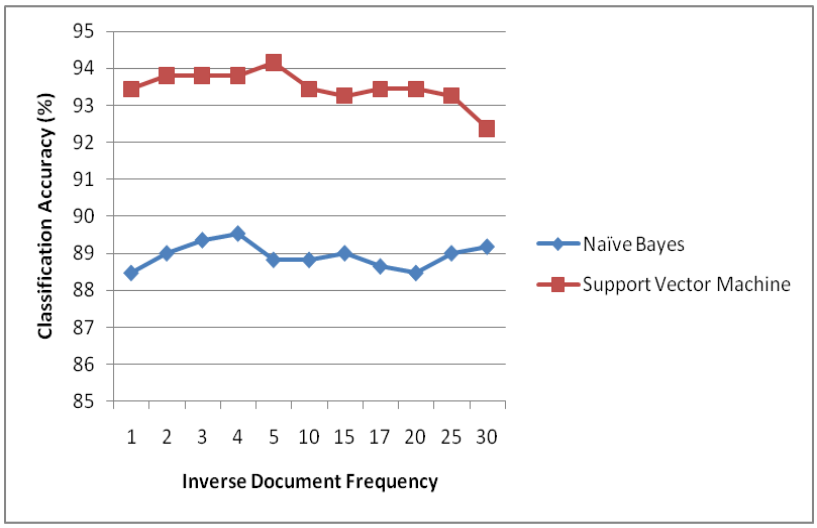

Fig 2. Inverse Document Frequency vs. Classification Accuracy

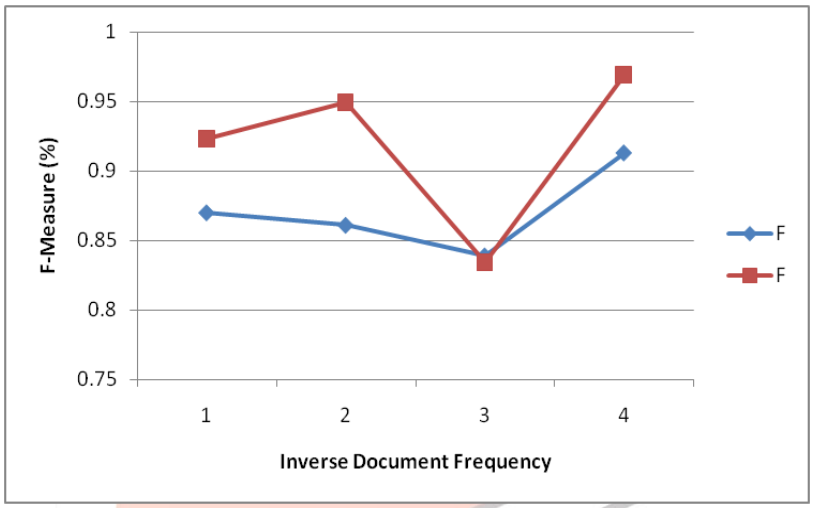

Fig 3. Inverse Document Frequency vs. F-measure

VII. CONCLUSIONS

Web page classification with html tag and inverse document within each tag - combination as features classifies web pages more accurately. Naive Bayes classifier performs the classification with 89.54 % accuracy, while Support Vector Machine performs classification with 93.79%. Thus it is concluded that Support Vector Machine is good for web page classification with combination of html tag and Inverse Document Frequency as feature. This work classifies web pages of four categories course homepages, faculty homepages, project pages and student homepages. Thus it can be used for developing educational universities' catalogs. This catalogs 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 course, faculty, project and students, which might be a tedious task otherwise as it requires going through all information. Also building such catalogs manually require lot of human effort.

VIII. FUTURE WORK

In this dissertation work, the web pages are classified into four categories, course, faculty, project and student. Thus it can be used for developing, maintaining Educational Universities' catalogs, which can be expanded by including more categories such as, Department, Event, staff etc. As this work is related to classification of educational web pages, analyzing them specifies the fact that most web pages contains copyrights and various images as noise. These are usually specified using or <font> and <img> tags, respectively. As these tags are not considered for feature extraction, 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 Feature selection method which has been used for this dissertation work is Inverse Document Frequency. Other methods such as Term Occurrence number, Term frequency (TF), Document frequency (DF) can be used.

ACKNOWLEDGEMENT

I would like to acknowledge 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] Xiaoguang Qi and Brian D. Davison, “Web Page Classification: Features and Algorithms ", ACM Computing Surveys, Vol. 41, No.2, Article 12, 2009.

[2] Gabriel Fiol-Roig, Margaret Miro-Julia, Eduardo Herraiz,"Data Mining Techniques for Web Page Classification” , 2011.

[3] 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.

[4] Porter, An algorithm for suffix stripping. Program, 14(3), 1980, 130-137.

[5] Jiawei Han, Micheline Kamber and Jian Pei, Third Edition, Data Mining Concepts and Techniques.

[6] Arul Prakash Asirvatham, Kranthi Kumar. Ravi," Web page classification based on document structure", Awarded second prize in National Level Student Paper Contest conducted by IEEE India Council, 2001.

[7] Myungsook Klassen, Chenxiao Wang,"Search Web Page Classification Using Form Structural Characteristics", 24th International Conference on Computers and Their Applications in Industry and Engineering (CAINE), 2011.

[8] Chih-Ming Chen, Hahn-Ming Lee, Yu-Jung Chang, "Two novel feature selection approaches for Web page classification", Expert Systems with Applications 36, 2009, 260-272.

[9] Zakaria Suliman Zubi, "Using Some Web Content Mining Techniques for Arabic Text Classification", Recent Advances on Data Networks, Communications, Computers, ISSN: 1790-5109.

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

[11] Victor Fresno, Raquel Martinez, Soto Montalvo, Arantza Casillas,"Naive Bayes Web Page Classification with HTML Mark-Up Enrichment", In proceeding of: Computing in the Global Information Technology, 0-7695-2629-2/06, 2006 IEEE.

[12] M.Indra Devi, Dr.R.Rajaram, K.Selvakuberan, “Automatic Web Page Classification by combining Feature Selection Techniques and Lazy Learners”, IEEE 2007.

[13] https://www.wiwi.hu-berlin.de/professuren-en/quantitativ/wi/forschung-en/dwm/standardseite-en.

[14] Tamanna Bhatia, “Link Analysis Algorithms For Web Mining", International Journal os Computer Science and Technology, issue-2, June-2011.

[15] http://searchsoa.techtarget.com/definition/stop-word.

[16] Jonathan Elsas, "HTML Tag Based Metrics for use in Web Page Type Classification", 2004.

[17] M. A. Shah, Dr. S. M. Deshpande, "Web Page Classification Based on Document Structure without Negative Examples", International Journal Of Computer Science And Applications Vol. 1, No. 1, June 2008

[18] Zakaria Suliman Zubi, "Using Some Web Content Mining Techniques for Arabic Text Classification", Recent Advances on Data Networks, Communications, Computers, ISSN: 1790-5109.

[19] Deng Cai, Shipeng Yu, Ji-Rong Wen, Wei-Ying Ma, "VIPS: A VIsion based Page Segmentation Algorithm",July-2004.

[20] Deng Cai, Shipeng Yu, Ji-Rong Wen, WeiYing Ma,"Improving Pseudo- Relevance Feedback in Web Information Retrieval Using Web Page Segmentation", 2003.