ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)
International Conference on Recent Trends in Computing and Communication (ICRTCC 2015)
Cochin College of Engineering & Technology
Vol. 4, Special Issue 1, June 2015

Web Page Categorization with Extended TDW Scheme

Arun P R[1], Sumesh M S[2], Eldhose P Sim[3]

PG Scholar, Computer Science and Engineering, Adi Shankara Institute of Technology and Science, Kalady, India[1]

Assistant Professor, Information technology, Adi Shankara Institute of Technology and Science, Kalady, India[2]

Assistant Professor, Computer Science and Engineering, Cochin College of Engineering, Malappuram, India[3]

Abstract: The exponential growth of internet over the past decade has increased millions of web pages published on every subject. Internet provides only a medium for communication between the computer and for accessing online document over this network but not to organize this large amount of data. There are different subject based web directories like Open Directory Project's (ODP) Directory Mozilla (DMOZ), Yahoo etc., these directories organize web pages in hierarchy. Due to the rapid growth of web pages the categorization demands the need of machine learning technique to automatically maintain the web page directory service. To assign a web page into a class the textual information in the page serves as a hint. Here we propose a method which uses an extended TDW scheme for feature representation and a naïve Bayesian to build the classification model. The web page categorization provides a wide range of advantages that ranges from knowledgebase construction, to improve the quality of web results, web content filtering, focused crawling etc.

Keywords: Categorization, Extended TDW Matrix, Naive Bayesian, Feature selection

I. INTRODUCTION

Web Mining is the branch of data mining which deals with the study of World Wide Web [1]. It refers to the use of data mining techniques to automatically find out and mine information from World Wide Web documents and services. In every second millions of bytes are added all over the world. As the number of web pagesare increasing the task of classification not feasible by manually which demands the need of a programmed technique? Web pages contain many irrelevant words than relevant words from advertisement, banner and special characters. It may also contain audio, image and/or video information associated with them. The web pages does not keep any standard formatting or size, pages are volatile in nature. Some web pages may be unstructured, some pages may be semi structured and some pages may be structured. To find the category of the web page, the textual information in the web page serves as a hint.

People are highly depending on the search engines for finding the required information. The results returned by a query always shown in a single rank based mechanism, by using the categorization the returned result can be differentiated based on the category as well as ranking in each category which can effectively improve result quality.The web page categorization method should address the synonym detection, rank based outcome, dimensionality of document representation, and existence of local noise, html tag based content representation and URL in the page. Instinctively the web page categorization is the task of classifying a web document under a predefined category. More formally, Let {W_1, W2... Wn} be a set of web page features and {C_1, C2…C_t} be the predefines classes. The task of classification assign a web page W_i into a class C_j based on its features. The features of a web page can obtain from the page itself and its neighbor page. The neighbor page feature may lead to false classification due to the existence of local noise. The proposed methods mainly consist of two phases feature extraction phase and classification phase. The rest of the paper is organized as follows. Section II reviews previous work on the machine learning and classification. Section III and IV discusses the classification of web pages and Naïve Bayes Theorem respectively, Section V presents our approach of classifying websites based on home pages using NB technique. Section VI discusses the results of our experiment. The last section summarizes the paper and gives some directions for future research.

II. RELATED WORKS

Most of the existing methods based on term frequency weighting scheme for feature representation. The term relevance of a class is obtained based on how many times a word present in the class, as before term frequency approach Boolean model where used. It provides a binary weighting scheme. This model has extended by Gerard Salton et al.[2]. .While considering a web page the importance of a word is varying based on where it belongs. Consider an example, two web pages shares a common term on of the page belongs to class Aand other in class B. In the first web page the term present in the title or header tag and in the second page it belongs to the paragraph tag. In the term document weighingscheme [3] the relevance of the term is higher for class A as compare

DOI 10.17148/IJARCCE

to class B.For the classification of web page there are different classifiers. Support Vector Machines (SVM) is a relatively new learning approach introduced by Vapnik in 1995 for solving two-class pattern recognition problems [4][5]. The method is defined over a vector space problem is to find a decision surface that best separates the data points in two classes. In order to define the best separation, we need to introduce a “margin” between two classes.

KNN stands for k-nearest neighbor classification, a well-known statistical approach which has been intensively studied in pattern recognition for over four decades [6] [7]. The KNN algorithm is quite simple: given a test document, the system finds the k nearest neighbors among the training documents, and uses the categories of the k neighbors to weight the category candidates. The similarity score of each neighbor document to the test document is used as the weight of the categories of the neighbor document.

LLSF stands for Linear Least Squares Fit, a mapping approach developed by Yang [8]. A multivariate regression model is automatically learned from a training set of documents and their categories. The training data are represented in the form of input/output vector pairs where the input vector is a document in the conventional vector space model (consisting of words with weights), and output vector consists of categories (with binary weights) of the corresponding document.

Neural network (NNet) techniques have been intensively studied in Artificial Intelligence. NNet approaches to text categorization were evaluated on the Reuters-21450 corpus by Wiener et al. [9]. Wiener et al. tried both a perceptron approach (without a hidden layer) and three-layered neural networks (with a hidden layer). The systems use a separate neural network per category, learning a non-linear mapping from input words (or more complex features such as singular vectors of a document space) to a category.

Naive Bayes (NB) probabilistic classifiers are commonly studied in machine learning [10] [11] [12]. The basic idea in NB approaches is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. The naive part of NB methods is the assumption of word independence, the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category. This assumption makes the computation of the NB classifiers far more efficient.

For the huge amount of training data SVM, KNN shows higher performance but SVM is very slow while training and KNN slow at testing time which again places an importance to the naïve Bayesian classifier.

III. PROPOSED METHOD

The proposed approach uses the extended term document weight (ETDW) scheme. Web page content is completely different from an ordinary text file so the relevance of a term is varying not only based upon the frequency of term appeared but also where it present in the document.The proposed method consist of three phases feature extraction, Feature selection and classification. The feature extraction stage takes an input web page. First step is to extract the web page content based on the html tags, then perform stop word removal. The stop words are consider to be the connectives and preposition in English language. For stop word removal we use the Google provided list of words. The resulting string content is tokenized.

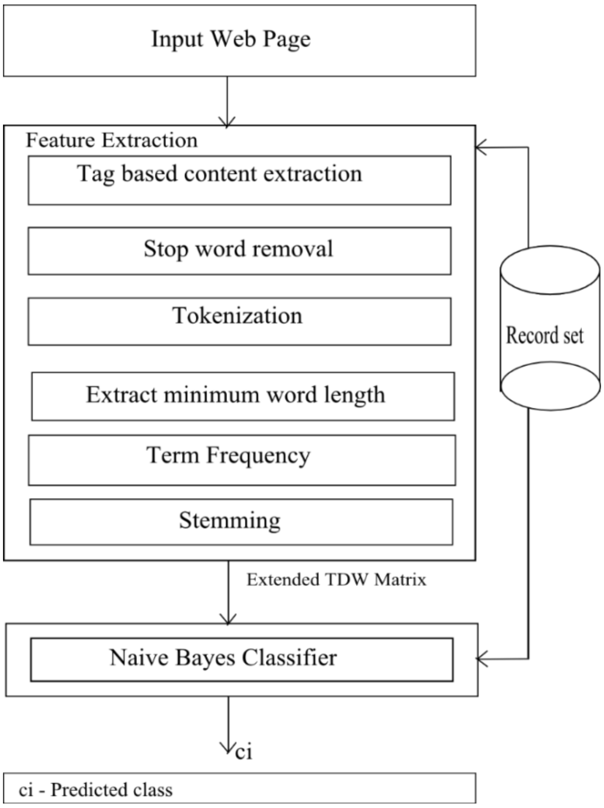


Fig 1: Proposed Architecture

The word mapping phase receives the tokenized content and for each word having a synonym extract the row of its synonyms and returns the minimum length word for stemming. For example a word x3 is having a synonym then extract the first synonym field value for x1 from the particular row, the first field will be minimum length synonym term for x3.

TABLE I. SYNONYM TERMS

Id	Synonym1	Synonym2	Synonym3
1	x1	x2	x3
2	y1	y2	y3

In many cases, morphological alternatives of words have similar semantic interpretations and can be considered as equivalent for the purpose of many applications. Porter’s stemming algorithm [13] is used to map the term into its root form. In order to reduce the dimensionality of the dataset we use a term frequency feature selection scheme.Compute the Extended TDW matrix for the web page based on the term appearing frequency and tag weighting. Letrec1, rec2, rec3 be three records. rec1= {t2, t1, t3} rec2= {t4, t1, t3} rec3={t2, t4, t1, t3}.The notation fti be the number of times the term appear in a specific tag. The TDW matrix with sentence feature is given in Table 2.

TABLE II. EXTENDED TDW MATRIX WITH SENTENCE FEATURE

Records	t1	t2	t3	t4	Class
rec1	$\sum ft1*w_t$	$\sum ft2*w_t$	$\sum ft3*w_t$	0	Ci
rec2	$\sum ft1*w_t$	0	$\sum ft3*w_t$	$\sum ft4*w_t$	1. 2. Cj3.
rec3	$\sum ft1*w_t$	$\sum ft2*w_t$	$\sum ft3*w_t$	$\sum ft4*w_t$	ck

w_t represent the tag weighting scheme. Here we use the weighting scheme [14].

TABLE III. WEIGHTING SCHEME

Term Field	Weight
URL	2
Heading	2
Title	2
Link to same website	1
link into another web site	0.5
Keyword	3
Description	3
Main Block	1

For the web page classification we use the naïve Bayesian classifier. The extended TDW matrix contains web page feature terms tagged with the tag-frequency weight. The naïve Bayesian is based on the bayes rule, it requires a simple representation of the document feature as a bag of words. Let R={r_1,r_2,…,r_n} be set of records and C={c_1,c_2,...,c_m} be the set of classes. Each record is labeled with the associated classes (r_i,c_j) in the training set. The probability of a document r_i being in class c_i using Bayes theorem is given by:

$$\boldsymbol{P(Ci \mid ri)} = \underset{c_i \in C}{\operatorname{argmax}} \frac{P(r_i \mid c_i)P(c_i)}{P(r_i)} \quad)$$

As P(r_i) is independent of the class, it can be ignored.Naïve Bayesian has an assumption that every attributes are consider to be an independent to each other.

$$P(ri|\ ci)= P(t_1 \mid ci)P(t_2 \mid ci)P(t_3 \mid ci)...P(t_n|\ ci) \qquad (2)$$

The probability of a term to a class can be computed by

$$P(t_j \mid c_i) = \frac{count(t_j, c_i) + 1}{count(c_i) + |V|} \qquad (3)$$

By equation (1),

$$\boldsymbol{P(Ci \mid ri)} = \underset{c_i \in C}{\operatorname{argmax}} P(c_i) \prod_{0<k<n+1} P(t_k \mid c_i) \qquad (4)$$

A web page is assigned to a class based on the posterior probability P(ci| ri).

IV. PROPOSED ALGORITHM

```
A.Algorithm :   Web_Page_Categorization
Input: Input Web page W,Record set R_i
Output:Class label C
ETDW  = Feature_extraction (W);
C = Classifier(ETDW,ETDW (Ri));
Return C;
```

```
B.Algorithm:                    Feature_extraction
Input: Input Web page W
Output:Extended_TDW_Matrix
1. E_content=tag_based_content(W);
2. Content=stop_word_removal(E_content);
3. T[]=Tokenize(Content);
4. For j=0 to |T|
5      Wm_c=WordMapping(T[j]);
6ETDW(W,Wm_c)=ETDW(W,Wm_c)+(count(Wm_c)*tag_weight);
7. End;
8. Return ETDW;
```

```
C:Algorithm                     WordMapping
Input: Tokenized term Ti
Output: Stemmed term
1. If (exist (Ti).synonym)
2. Extract_synonym [] =Synonym (Ti);
3. Ti=Minimum length (Extract_synonym[]);
4. End
5. Return porter_stem(Ti);
```

```
D:Algorithm              Classifier
Input: ETDW, ETDW (Ri)
Output: Predicted class C
 1.Foreach Class C
 2.  Posterior_prob [] = Compute (P (Ci | ETDW))
 3.   Return Max (posterior_prob (Ci));
```

V. EXPERIEMENT AND RESULT

A. Data Collection

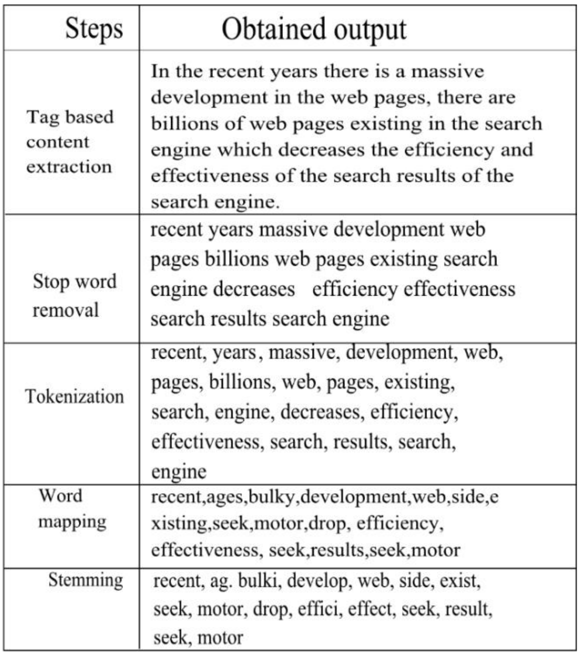
The data is collected based on different categories. In order to create the dataset, services of various search engines and subject directories were used. Search engines like goggle, bing etc, and DMOZ subject directory has used. Created a repository of web pages obtained by querying some specific keywords and collected all similar web pages with respect to the higher rank position. Some of the results were omitted due to the lack of required contents retrieved and based on required file formats. Each page thus obtained are processed to extract the features which weighted according to the weighting scheme and created the extended term document weight matrix with a class attribute for the training data.

B. Experimental Setup and Result

To conduct the required experiment, we created an online tool which is capable of extracting features form web page either by giving a URL or by upload the web page from the local system. The system build an Extended TDW matrix for the input web pages by applying stop word removal, stemming and word mapping. During the training phase web page given as input with the class label.

Table 4 shows a sample feature extraction. Based on the stemmed output the extended TDW matrix is formed. The matrix formed by arranging the term as attribute names. The weights are inserted into the column with respect to the record entry.

TABLE IV. PREPROCESSING



Steps	Obtained output
Tag based content extraction	In the recent years there is a massive development in the web pages, there are billions of web pages existing in the search engine which decreases the efficiency and effectiveness of the search results of the search engine.
Stop word removal	recent years massive development web pages billions web pages existing search engine decreases efficiency effectiveness search results search engine
Tokenization	recent, years, massive, development, web, pages, billions, web, pages, existing, search, engine, decreases, efficiency, effectiveness, search, results, search, engine
Word mapping	recent,ages,bulky,development,web,side,existing,seek,motor,drop, efficiency, effectiveness, seek,results,seek,motor
Stemming	recent, ag. bulki, develop, web, side, exist, seek, motor, drop, effici, effect, seek, result, seek, motor

C. Result Analysis

For evaluating the degree of accuracy, efficiency and scalability of our proposed approach, we have used repository that contains the web pages documents obtained through querying in Google search engine. The performance of the proposed approach is evaluated with the help of evaluation metrics such as, Precision, Recall.

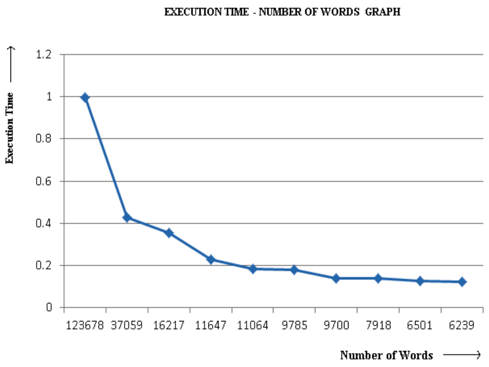
$$\text{Precision (P)} = \frac{\text{Number of relevant web pages retrieved}}{\text{Total number of web pages retrieved}} \quad (5)$$

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

TABLE V. PERFORMANCE MEASURE

Category	Precision %	Recall %
Education	90.9	89.81
Banking	89.7	89.05
Tourism	91.1	90.73
Cooking	89.55	89.44
Hospitals	89.46	89.31
Beauty and Health	91.29	90.92
Program documentation	92.97	92.61
Sports	93.56	92.47
Entertainment	91.01	90.78
Accuracy	91.06	90.56

While creating a repository of 100 pages, a TDW matrix of almost size 100 x 1700 is created in preprocessing phase. When the word mapping scheme is applied the term count reduced to almost 1400.



The proposed feature weighting scheme is very efficient, but the time required for the web page parsing increases depending upon the word count increases. The execution time can be reduced by dividing the computation into smaller task and perform it in parallel.

VI. CONCLUSION AND FUTURE WORKS

The web page categorization is stance as one of the main threat to the huge increasing of web pages in heterogeneous sources, which demands an efficient machine learning technique to categorize the web page. There are various methods have proposed with different classifier. In the proposed approach we use an extended term document weighting scheme (ETDW) for feature representation and naïve Bayesian to build the classifier model. According to this weighting the experiment result shows that the proposed method is efficient and improved the precision and recall.

We believe that this work represents an important step toward the web page categorization. The proposed method can be further improved by implementing a parallel execution framework for feature extraction.

ACKNOWLEDGMENT

The authors thank the authorities of their working organization and management for their support and the encouragement to pursue research in the chosen field of study.

REFERENCES

[1] Andrei Z. Broder, Steven C. Glassman, Mark S.Manasse, and Geoffrey Zweig, "Syntactic Clustering of the Web", In Proceedings of the Sixth International.
[2] Gerard Salton,EdwardA.Fox,Harry WU, "Extended Boolean Information Retrieval", Communications of the ACM 1983 Volume 26 Number 12
[3] MidhunMathew,Shine N Das,TR Lakshmi "A Novel Approach for Near-Duplicate Detection of Web Pages using TDW Matrix" IJCA 2011
[4] V. Vapnic. "The Nature of Statistical Learning Theory", Springer, New York, 1995.
[5] C. Cortes and V. Vapnik. "Support vector networks". Machine Learning, 20:273{297, 1995.
[6] Belur V. Dasarathy. "Nearest Neighbor (NN) Norms: NN Pattern Classifcation Techniques". McGraw-Hill Computer Science Series. IEEE Computer Society Press, Las Alamitos,California, 1991.
[7] B. Masand, G. Lino_, and D.Waltz. "Classifying news stories using memory based reasoning". In 15th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92), pages 59{64, 1992.
[8] Y. Yang and C.G. Chute. "An example-based mapping method for text categorization and retrieval". ACM Transaction on Information Systems (TOIS), 12(3):252{277, 1994.
[9] E. Wiener, J.O. Pedersen, and A.S. Weigend. "A neural network approach to topic spotting". In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), 1995.
[10] Tom Mitchell. "Machine Learning". McGraw Hill, 1996.
[11] D.D. Lewis and M. Ringuette. "Comparison of two learning algorithms for text categorization". In Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), 1994.
[12] I. Moulinier. "Is learning bias an issue on the text categorization problem?", In Technical report, LAFORIA-LIP6, Universite Paris VI, 1997.
[13] M.F. Porter, "An algorithm for suffix stripping", Program, Vo.14, no.3, pp. 130-137, Jul. 1980.
[14] Shine N Das, Midhun Mathew, PramodK.Vijayaraghavan, "An Approach for Optimal Feature Subset Selection using a New Term Weighting Scheme and Mutual Information", Proceeding of the International Conference on Advanced Science, Engineering and Information Technology, Malaysia, 2011, pp 273-278, January 2011.

BIOGRAPHIES

Mr. Arun P R, PG Scholar in Adi Shankara Institute of Engineering and Technology, specialization in CSE under MGU. B.Tech. degree in Computer Science & Engineering from MGU, India, in 2013.

Mr. Sumesh M S, presently working as an Assistant professor in Department of Information Technology, Adi Shankara Institute of Engineering and Technology. Completed M. Tech (CSE) from CUSAT and B. Tech (CSE) degree from MGUCE. He has four year experience in teaching and five year experience in industrial.

Mr. Eldhose P Sim, Assistant Professor, Computer Science and Engineering, Cochin College of Engineering, Malappuram, India. B. Tech. degree in Computer Science & Engineering from MGU and M. Tech from Anna University.