JCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

An Overview of Various Classification Concepts of Web Page Content

R Khan[1*], R K Gupta[2], V. Namdeo[3]

[1,2,3]Department of Computer Science, SRK University, Bhopal, India

[*]*Corresponding Author: rkresearch2019@gmail.com, Tel.: 9191537831*

Available online at: www.ijcseonline.org

Accepted: 24/Jan/2019, Published: 31/Jan/2019

Abstract**—** This paper collects the information about contents available over webpage since the Web is a huge stock of information that requires precise automated classifiers for web pages to manage web directories and increase search engine performance. In the Web page classification problem, each term can be used as a feature of each HTML / XML tag of each web page. This is an efficient way to select the best features to reduce the functional space of the derived Web page classification problem here. Content classification of web pages is essential for many Web information retrieval tasks, such as web directory management and targeted scanning. The uncontrolled nature of web content poses additional problems for the classification of web pages over traditional text classification. However, the interdependent nature of hypertext also provides functions that support the process. As with the work described in the Web page classification, the meaning of these Web-specific functions and algorithms describes leading practices and follows the assumptions underlying the use of adjacent page information.

Keywords**—** Algorithm, Assumption, Classification, Directory, Features, Information, Process, XML, Wepage etc**.**

I. INTRODUCTION

As the popularity of the band increases, the amount of information on the Web has also increased. This growth of information has led to the need for accurate and rapid classification of Web pages to improve search engine performance. Automatic classification of the website is a supervised learning problem in which a set of web documents tagged for training a classifier, then the classifier is used to assign one or more predefined category labels web pages for future use [1]. Automatic classification of the website is not only used to improve the performance of search engines, it is also essential for the development of web directories, discussion of specific topics Web, contextual advertising links on the analysis of the structure current site and to improve the quality of web search. Several methods of classification such as decision trees, Bayesian classifier, support vector machines, k-nearest neighbors were developed. Web page classification is developed based on technique of text classification [2-3]. Text representation and extraction of feature are essential questions of both text mining and information retrieval in the course of text classification which indicates text information using quantified feature words from texts. An improved formula calculating mutual information is proposed to describe the structure of web pages accurately [4]. Some optimization types shown in figure 1.

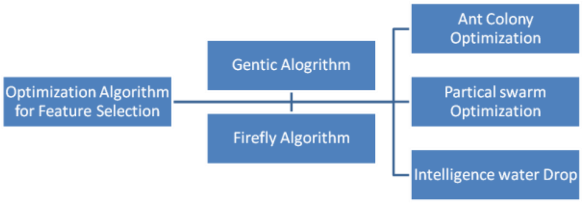

Figure 1: Types of Optimization Algorithm

II. RELATED WORK

[1] In this paper Author proposed a methodology to classify web pages hierarchically in order to achieve topic-wise modeling of websites using multi label tree classifier, a variant of classification where instances may belong to multiple classes at the same time. Data from an implementation of multi label tree classifier shows marked improvements in processing multi-class classification in comparison to conventional hierarchical classification techniques.

[2] In this paper Author survey, the widely used algorithms for text classification are given with a comparison of the recent researches in classification field for Arabic and English languages to conclude which is the best algorithm that we can apply for both Arabic and English Languages.

[3] In this paper Author introduced the details of a Chinese Web-page classification system that we implemented. Experiments show that our web-page pre-processing and feature selection method is effective. The classification accuracy acquired on a Chinese Web-page dataset is satisfying.

[4] In this paper Author present a novel sensitive information classification algorithm and topic tracking algorithm for Web pages contents. First, a text sensitive information classification method is proposed based on a vector space model and cosine theorem. Experiments show that the classification of the text sensitive information is very effective and result of topic tracking is ideal.

[5] In this paper Author discussed about multiple classifiers are built, one for each training domain, and the block classification proceeds through combining them.

[6] In this paper Author placed the topic knowledge repository is built to find the semantic relation between Web pages, and the similar relation and the associated relation are defined to describe the semantic relation, which helps to provide knowledge service for user and other services.

[7] In this paper Author studies the process and methods of text classification. Based on Naive Bayesian algorithm and the semi-structured feature in Web page information, this paper proposes an improved Algorithm for Web page text Information classification which utilizes Html tag Information in classification. Experiments show that this algorithm is feasible and effective and can apply to information extraction in topic search engine.

[8] In this paper Author tried to use a novel methodology to classify these documents. The approach that we have used is a combination of text classification and other binary classification. Using this we have been able to get an overall accuracy of 88%.

[9] In this paper Author discussed about A classification approach for Tibetan web pages is introduced in this paper. It takes advantage of the class feature dictionary and Rocchio classification algorithm to classify the Tibetan web pages into the predefined classes rapidly and accurately. The experimental results present that the approach has better classification accuracy for Tibetan web pages classification. It is useful and helpful for the construction of the statistical and rule-based classification of Tibetan texts as well as construction of high-quality Tibetan corpus.

[10] In this paper Authors shows that the accuracy of webpage classifiers can be improved by extracting meaningful strings with an unsupervised clustering method.

III. EXISTING METHODOLOGY & TOOLS

Web page categorization becomes a key technology in the transformation and organization of a mass of documents and data. The function is selected to improve the processing technology text hyperlink factor that believes in the maximum entropy model. Experiment found that the process is more efficient [5].

Ranking web pages is the technology developed in the form of text classification. The main text classification and classification of web pages difference is that websites have a lot of other information such as text links sound image, etc., which are very important in classification. So is important that we have text combined with information from the websites of web analysis [6-8]. There are so many classifiers exist to efficiently classify the contents as given in Table 1 below:

Table 1: Types of Classifier

Classifier	Description
Bayes linear	Assumes Gaussian distribution of features with equal covariance matrices for each class.
kn-Nearest neighbor	A robust non-parametric classifier. Classification has high computational complexity.
Neural network	The multi-layer perceptron (a non-parametric classifier) is the standard network to use for supervised learning.
Decision tree	Non-metric method. Gives a set of rules that can be understood.
Nb	Models that assign class labels to problem instances, represented as vectors of feature values where the class labels are drawn from some finite set.

WEKA

Weka, formally called Waikato environment for knowledge learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains [3,7 & 9]. weka supports many different standard data mining tasks such as data pre-processing, classification, clustering, regression, visualization and feature selection. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. weka is an open source application that is freely available under the gnu general public license agreement. Originally written in c the weka application has been completely rewritten in java and is compatible with almost every computing platform.

J48 CLASSIFIER

Classification is the process of building a model of classes from a set of records that contain class labels [5,10 & 12]. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found.

MATLAB

MATLAB development began at the end of 1970. It was designed to give student's access to LINPACK and EISPACK without learning Fortran. It soon spread to other universities and found a strong audience within the applied mathematics community. Jack Little, an engineer, was exposed during a visit Moler made to Stanford University in 1983. Recognizing its commercial potential, he joined with Moler and Steve Bangert. Rewrote MATLAB in C and founded in 1984, The MathWorks for further development [3,13 & 14].

NB CLASSIFIER

Naive Bayes has been extensively studied since the 1950s has been presented with a different name in the community of the search text in the 1960s [1]: 488 and still a (basic) popular method of categorization of text, judging issue documents as belonging to a particular category (such as emails or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-treatment, it is competitive in this field and more advanced methods, including support vector machines also finds application in automatic medical diagnosis. Naive Bayes classifiers are highly scalable, which requires a number of linear parameters in the number of variables (features / predictors) with a learning disability.

IV. PERFORMANCE METRICS

1. Precision: The term precision means two or more values of the measurements are closed to each other. The value of precision differs because of the observational error. The precision is used for finding the consistency or reproducibility of the measurement [12].
2. Recall : Recall is the ratio of correctly predicted positive observations to the all observations in actual class
3. Accuracy: Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same [9-13].
4. F-Measure: is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.

V. CONCLUSION AND FUTURE SCOPE

The increase in the amount of information on the Web has caused the need for precise automated classifiers for the Web pages to keep web directories and increase the search engine performance. Every tag and every term on every web page possible to be considered a characteristic, effective method are necessary select the best features to reduce the functionality of the web page classification problem. Search for Web classification with as for its characteristics and algorithms, we I conclude this summarize the lessons learned from existing research and highlighting future opportunities in web classification. Classification activities include the assignment of documents on the basis of subject, function, feeling, kindness, etc. I studied number of web page classification techniques, but due to the rapid growth of data on the Internet, there is still need for efficiency technique.

REFERENCES

[1] Daya Gupta ; Harsh Tripathi ; Mayukh Maitra, Classifying web hierarchically using multi label tree classifier, 2015 Annual IEEE India Conference (INDICON), 2015

[2] Sumaia Mohammed Al-Ghuribi ; Saleh Alshomrani, A Simple Study of Webpage Text Classification Algorithms for Arabic and English Languages, 2013 International Conference on IT Convergence and Security (ICITCS), 2013

[3] Chinese Web-page Classification Study, Weitong Huang ; LuXiongXu ; Junfeng Duan ; Yuchang Lu, Chinese Web-page Classification Study, 2007 IEEE International Conference on Control and Automation, 2007

[4] Guixian Xu ; Ziheng Yu ; Qi Qi, Efficient Sensitive Information Classification and Topic Tracking Based on Tibetan WebPages,IEEE Access, 2018

[5] Jinbeom Kang ; Joongmin Choi, Block Classification of a Web Page by Using a Combination of Multiple Classifiers, 2008 Fourth International Conference on Networked Computing and Advanced Information Management,2008Sara Chadli,Mohamed Emharraf and Mohammed Saber "The design of an IDS architecture for MANET based on multi-agent" International Colloquium on Information Science and Technology (CiSt),IEEE,2014

[6] Feiyue Ye ; Zhian Yu, Finding the Semantic Relation between Web Pages through Topic Knowledge Repository, 2009 Ninth IEEE International Conference on Computer and Information Technology, 2009

[7] He Youquan ; Xie Jianfang ; Xu Cheng, An improved Naive Bayesian algorithm for Web page text classification, 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011

[8] Moonis Javed ; Aly Akhtar ; Akif Khan Yusufzai, Classification of Web Pages as Evergreen Or Ephemeral Based on Content, 2015 International Conference on Computational Intelligence and Communication Networks (CICN), 2015

[9] Guixian Xu ; Chuncheng Xiang ; Xu Gao ; Xiaobing Zhao ; Guosheng Yang, Automatic Classification of Tibetan Web Pages,

International Conference on Computer Science and Electronics Engineering, 2012

[10] Jie Chen, Jian Li, Hao Liao, Qingsheng Yuan, Xiuguo Bao; Study on Meaningful String Extraction Algorithm for Improving Webpage Classification, IEEE, 2011

[11] Prabhjot Kaur ,Web Content Classification: A Survey, IJCTT, 2014

[12] Sankalap Arora,Satvir Singh, The Firefly Optimization Algorithm: Convergence Analysis and Parameter Selection, IJCA, 2013

[13] Bundit Manaskasemsak and Arnon Rungsawang, Web Spam Detection using Link-based Ant Colony Optimization Apichat Taweesiriwate, IEEE, 2012

[14] Ontological Based Webpage ClassificationWui Kheun Ong,Jer Lang Hong,Fariza Fauzi,Ee Xion Tan, IEEE, 2012

Authors Profile

Mr.R.Khan is pursuing Master of Technology in Computer Science & Engineering Department of Sarvepalli Radhakrishnan University, Bhopal (MP). He focuses on the field of computer security to strengthen computer security learning and importance.

Mr.R.K.Gupta is a well known Professor of Computer Science & Engineering Department of Sarvepalli Rdhakrishnan University, Bhopal (MP). He is pursuing Ph.D. in Computer Science from Barkatullah University Bhopal. His research interest and specialization falls upon intrusion Detection System and various Computer Security Strategies.

Dr.V.Namdeo Head of Computer Science & Engineering Department of Sarvepalli Radhakrishnan of Sarvepalli Rdhakrishnan University, Bhopal (MP). She is equiped with huge knowledge and experience of various fields of Computer Science fields.