A Three Stage Classifier for Efficient Website Categorization

Dhanashri S. Hulavale
ME - Computer Engineering
DYPCOE, Ambi
Pune University

Saurabh H Deshmukh, PhD
Department of Computer Engineering
DYPCOE, Ambi
Pune University

ABSTRACT
Website categorization is one of the challenging tasks in the world of ever increasing web technologies. There are different way to categorization of web pages using different features and approach. Website contains lot of information like text, images, animation, video and links. So this information is call as features of website. For the website categorization purpose all Feature have most important role. The web has a lot of information in the form of images, video, animation and text etc present in the document. In proposed System uses number of feature of website and use three different classifier for website classification are naive bays classifier, linear classifier- perceptron and stochastic classifier. Here eight major categories of website have been selected for categorization; these are business & economy, job search, and science, education, sports, news & media, government, entertainment. Proposed system gives ranking to website. It will be more helpful for software developer or website designer for evolution of their site using our system so that they can judge that their website belongs to respective category or not.

Keywords
Linear classifier- perceptron, Machine learning, naive bayes, stochastic classifier

1. INTRODUCTION
This technique is for automatic categorization of the website into different classes or categories. Automatic categorization of web pages mostly based on similarity between documents contents or their structures [1][2][3].There are different of way to categories webpage using content of webpage like hyperlink , text and meta keyword [4] [8] [14]. A proposed system is to facilitate the user for categorization the website. Website are classify into eight major categories are entertainment, sports, news & media, job search, business & economy, education, government and science. Use the different features of website for categorization.
Ranking for given website for respective category so that the user can understand that the content of website or feature of website is belongs to that category or not. It will be more helpful for software developer or website designer for evolution of their site using our system so that they can judge that their website belongs to respective category and looks like. The main objective is to provide efficient way for categorization of website and helps the search engine to more efficiently classify the web pages. It is observed that a user use limited feature for website categorization it get poor result

2. LITERATURE REVIEW
Classification plays an important role in many information management and retrieval tasks. Web page Classification can also help improve the quality of Web search [3]. The web is a huge repository of information so there is a need for categorizing web documents and to facilitate the search and retrieval of pages. Existing algorithms mostly use text content of the web pages for classification. The web has a lot of information in the form of images, video, animation and text etc present in the document. [8] WebPages have many feature are buzzword, static &dynamic pages, animation, internal & external Links and Images etc.[9] For text feature needed to analyse the web content and metadata in relation. For text categorization Naive Bayes classifier gives best result but the problem of judging documents because a single word are belonging to one or more categories such as spam or legitimate, sports or politics, etc. [10] Linear classifier that can decide whether an input is related to one class or another. The stochastic classifier is used to define that the given website is how much percent from given category so calculate range and unit of distribution for that range. Stochastic classifier to will test that the score of given website are how much unit are satisfying the respective category.

3. PROPOSED SYSTEM
In proposed System use crawler is use to store website then using Feature Extraction extracts features of website like buzzword, Ration of internal & External Link, Images, animation, static & dynamic pages specialized keyword etc. Web pages contain many things which is considered as HTML TAG. The all contained in HTML file as tree structured. Tag in HTML file show that specific use and Significance, web pages are nothing but collection of this tags and show result of that as output which call web page which look in the browser thus to use that content for categorization purpose. These contents are: Links, media, static and dynamic pages, images etc..For developing a system for webpage categorization use feature extraction method, where features of website are extracted and analysis are done based on features.Use three different classifier are Naive Bayes classifiers, linear classifier-perceptron and stochastic classifier for website Classification. Website are classify in 8 different categories are entertainment, sports, news & media, job search, business & economy, education, government and science. Ranking for given website for respective category so that the user can understand that the content of website or features of website are belongs to that category or not. It will be more helpful for software developer or website designer for evolution of their site using our system so that they can judge website are belongs to that category or not.

It will be more helpful for software developer or website designer for evolution of their site using our system so that they can judge that their website belongs to respective category and looks like. Stochastic classifier which will gives result for ranking in unit of percentage. The stochastic classifier will use the unit number of features which belongs to respective category and also test that how many percent

that will match. Result of stochastic classifier show that the respective site are belongs to which category and how many percent. The website having higher percentage will have high ranking and website having low percentage will have low rank.

3.1 Classifiers

3.1.1 Naive Bayes classifiers

Naive Bayes is best for text categorization but many problems in text classification because one word belong to different category such as spam or legitimate, sports or politics, etc. Naive bayes will test all keywords and its respective category and score of those keywords are considered for classification of website. It will be tested for each keyword.

Word W1 will have category score in numbers like c1, c2,c3…c_n.

Y_c will be final score

C={c1,c2….c_n}

W={w1,w2,…w_n}

For each W

$$Yc = \sum_{1}^{n} scoreofW$$

$$ScoreofW = \sum ifscorethreshold \; C1 \dots Cn \leq \frac{}{\sum_1^n C1 + c2 \dots Cn}$$

If the score of given word for category are above threshold value then that word would be considered from that respective category otherwise it will be neglected. For flexibility purpose to keep threshold variable value which can be changed at runtime so that variable results are measured and accuracy of system will be flexible.

3.1.2 Linear classifier- perceptron

Proposed works have learning with supervision so that used this technique which will be less complex and flexible for testing purpose. The perceptron technique used to decide whether the given website is to be tested against stochastic classifier or not. It is like binary testing of website feature which are in specified range or not. If the website is in according to specific range then it will be forward consider for stochastic classifier testing. Feature extracted from site are stored in database as number of links, media, images and at the learning time summary of that also maintained. So there will be some upper limit and lower limit for each category are available in the form of numbers. At the testing time if website contents are within that limit or above lower limit then it would considered for further test.

U_l as upper limit;
L_l as Lower limit;

If Category L

$$< website\ L\ then \int_0^1 (1 \text{ for relevant and else } 0)$$

The perceptron will give binary result as the given site is relevant with given category or not.

3.1.3 Stochastic classifier

The stochastic classifier is used to define that the given website are how much percent from given category so calculate range and unit of distribution for that range. and maximum distribution is 100% and minimum distribution is 0.01%, by using stochastic classifier to will test that the score of given website are how much unit are satisfying the respective category.

$$RangeUnit(x) = \frac{Max - Min}{100}$$

In this case Max-Min will be always above zero or positive integer value because to consider that this test will performed after perceptron positive test.

Category Score =

$$\frac{c\text{urrentScore(for given paramater)} - \text{Min}}{unit(X)} * 100$$

So that it will result category score from 0 to 100, there may chance that there will result in some case above 100 but to will consider that as 100.

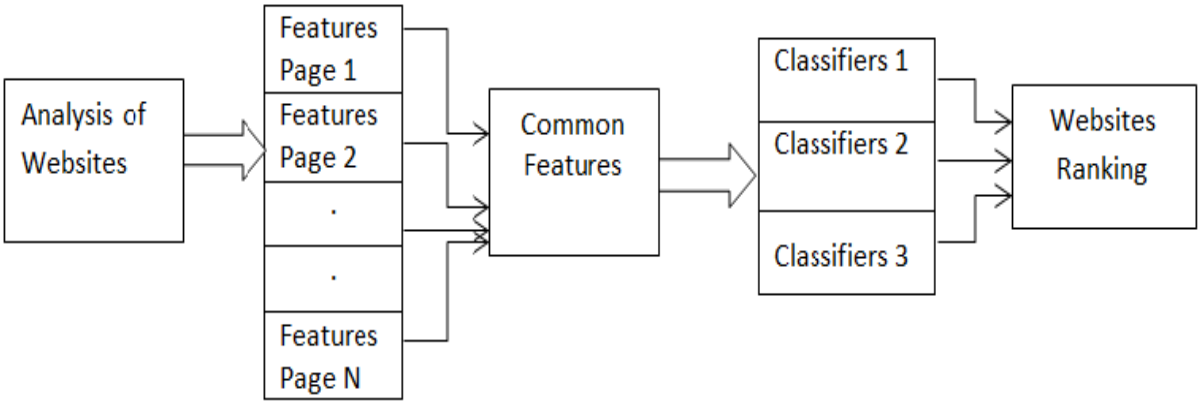

Fig 1: Proposed System Architecture

3.4 Website Ranking

Ranking for given website for respective category so that the user can understand that the content of website or features of website are belongs to that category or not. It will be more helpful for software developer or website designer for evolution of their site using our system so that they can judge that their website belongs to respective category and looks like. Stochastic classifier which will give result for ranking in unit of percentage. The stochastic classifier will use the unit number of features which belongs to respective category and also test that how many percent that will match. Result of stochastic classifier show that the respective site are belongs to which category and how many percent. The website having higher percentage will have high ranking and website having low percentage will have low rank.

4. CONCLUSION

This paper proposes a model for website Ranking Website Ranking will be more helpful for software developer or website designer to judge that their website belongs to respective category. Three different classifier are Naive Bayes classifiers, linear classifier-perceptron and stochastic classifier use for website Classification. Website are classify in 8 different categories and gives ranking to website. So it helps the search engine to more efficiently classify the web pages.

5. REFERENCES

[1] Arul Prakash, Kranti Kumar Ravi,Asirvhatam -Web Page Catego- rization based on Document Structure. In International Institute of IT, Hyderabad, India.

[2] J. B. Leela devi, 2dr. A. Sankar-IMPROVEMENTS IN NEURAL NETWORK FOR CLASSIFICATION OF WEB PAGES 1j. B. Leela devi, 2dr. A. Sankar 1university research scholar, anna university, Tamil Nadu, India 2. Associate Professor, PSG College of Tech- nology, Coimbatore, India.

[3] BRIAN D. DAVISON and XIAOGUANG QI. Web Page Classi- cation: Features and Algorithms. In ACM Computing Surveys, Article 12, Publication date: February 2009

[4] H.Yu, J.Han, and K.C.C.Chang.Pebl, -Positive example based learning for web page categorization. In Canada, KDD, Edmon- ton, Canada, 2002,

[5] In Iran University of Science Technology (IUST), Tehran, Iran.Arash Rezaei, and Behrouz Minaei-Bidgoli-Comparison be- tween the Classification Methods using Type of Attributes and Sample Size.

[6] Sundus Hassan and Muhammad Shahid in Computer Science Department NUCES-FAST, Karachi Campus,Comparing NB Clas- sifiers SVM for Text Classification.

[7] Tat-Seng Chua Hui Yang. Effectiveness of web page categoriza- tion on Finding List Answer, In National University of Singapore.

[8] E-H. S. Han, K. Hastings, D. Boley, M. Gini,V. Kumar, B. Mobasher, R. Gross, E-H. S. Han, J. Moore,K. Hastings and G. Karypis In 1999,DecisionSupportSystem, web document categorization using Partitioning-based clustering .

[9] G.W.F.S. Lawrence, W.P.Birmingham, A.Kruger , D.M.Pennock - Improving category specific web search by learning query modifications,san diego, Saint 2001, California.

[10] Arul Prakash, Kranti Kumar and Ravi,Asirvhatam -Web Page Categorization based on Document Structure. In International Institute of IT at Hyderabad.

[11] Komal Kumar,Pikakshi Manchanda and Sonali Gupta.The Auto- mated Classification of Web Pages Using Artificial Neural Net- work .Department of Computer Science, Faridabad, India