Automatic Classification of Web Pages using Back Propagation

Poonam Nagale
Student
DYPSOET,Lohegaon
Pune, India

Arti Waghmare
Asst.Professor, Department of Computer Engg.
DYPSOET,Lohegaon
Pune, India

ABSTRACT
Word Wide Web is huge repository of information. So there is tremendous growth in internet access by the user. To fulfill the requirement of numerous users, new sites are launched by the companies in the market every day. In previous we have to use conditional filtering to avoid illegal sites in campus. For that we have to manually provide that link to be blocked and also to keep proxy on such link, we have to pay for service. The manual categorization of links is possible as it is designed for humans only but it requires lots of man power also needs extra care while clustering. So, it's quite boring job. To overcome these problems we had designed a system which causes automatic categorization web links with the help of artificial neural networks (ANN).

Keywords
Feature extraction, Back propagation, incremental learning, ANN

1. INTRODUCTION
The rapid growth in a utilization of world wide web will require to invent new method or enhance the performance of existing system by removing its drawback to classify the web pages automatically into some classes or categories. This will not only benefited to user which are accessing internet and server but also the search engine for accessing requisite websites which will consumes less time and also improve the performance of system. We have made study of classification of web pages automatically, and understand similarity between document content and their structure are most preferred by most of the categorization technique [4] [5].

Most of the web page categorization techniques studied will be focus on certain features of web pages, which are not sufficient to categories the web page. There are some categorization techniques used which will classify the pages using the meta keyword, hyperlinks structures, document structure, and automatic text categorizations[6][7][8][9].

Artificial neural network is immerging trend in market. The simple features of ANN like it's architecture is like human brain. Once ANN is trained with sufficient and appropriate training it able to work on any data unseen before as per his training. ANN covered following features like Pattern Classification,Clustering/Categorization,Function Approximation, prediction/forecasting, Optimization, Content-address Management.

Pattern Classification: Pattern classification defined as a function that is used by two or more than two classes for the mapping of input feature space to an output space of that classes. Artificial Neural Networks (ANN) plays an effective role in the field of pattern classification, by using training and testing data for building a model.

Clustering/Categorization: In categorization process ideas and objects are recognized, differentiated ,and understood. In categorization objects are grouped into categories, to fulfil the some specific purpose. Ideally, a category will responsible for breaking a bonding between domain clustering using adaptive preprocessing the object and subject of knowledge. Clustering is works on unsupervised learning problem. Clustering is a process of arranging objects into a group having similar members in some way. Clustering is used for parallel processing, load balancing and fault tolerance.

Function Approximation: Function approximation is technique which is able to finds the underlying relationship from a given finite input-output data . Which is the major problem detected in majority of real world applications, such as prediction, pattern recognition, data mining and classification. Prediction/Forecasting: forecasting is used estimate a future values given past values. Prediction is used to calculate a value whether it is future, current or past with respect to the given data.

Optimization: Optimization function is used to maximizing or minimizing values in certain range of available options in some function relative to some set in a certain situation. This function gives the comparative result of various choices to provide result that might be best.
Content addressable memory: It is special type of computer memory which is useful in application who requires very high speed searching. So we are using ANN to classify the web pages.

Due to this features of ANN it is becoming popular in today's scenario. So, ANN will made easy clustering of web links into specified domain. As data increasing over time, that data should be updated or retain itself to give accurate result. The time require to mine data is about 90% and remaining 10% time require to process that data so, it is become necessary to preprocess data before it is feeded to predictor for further processing. To achieve this approach is to use incremental learning processing that is instance learning and batch processing. In instance learning updating old things with newer one keeping old things as it is. In batch processing system waits till the fixed data is available for the system operation The proposed system has used case sensitivity search , also ues standardization approach [2].

2. MATHEMATICAL MODEL
Let S be a set that describes the system.

S = {WS,WC,LN,U,N,A,D }Where,

Set Theory

1. WS= {L1, L2….. , Ln}

Ws is set of links of web sources and Li is the any http links of web sites.

2. WC= {WC1}

Wc is the set of web crawlers to retrieve information from web sources.

3. U={U1, U2,….,Un}

U is set of end users.

4. N={N1,N2,….,Nn}

N is the set of neurons.

5. LN={LN1,LN2,LN3}

LN is the set of layer of neurons.

6. D= {Dk, Db}

D is the set of databases where Dk is to retrieve keywords from links data & Db contain

classified links with it's domain.

7. A is the Admin which is unit set.

3. LITERATURE SURVEY

Inma Hernández , Carlos R. Rivero , David Ruiz , Rafael Corchuelo had proposed system in that they used url based classification. They had build number of URL patterns to represent the number of classes present in a web link. So, classification of web pages has did by coordinating web page's URL to the defined patterns. There proposal had satisfying the lightweight crawling by accessing header only, unsupervision and classifying without downloading [1].

Indre_ Zliobait e and Bogdan Gabrys had proposed a system which said that preprocessing of data is more important . as predictive model requires retaining and updated data to maintain the accuracy of the system. If only predictor is updated the system will failed to notice the changes. Secondly, If preprocessing is done repeatedly by scratching old data evry time then there is necessity of retraining of predictor. So, they had used adaptive approach to preprocess streaming data. And also addresses the problems of ADPP [2].

Rahul Isola and Amiya Kumar had proposed a system which predicts disease according to symptoms matched by performing iterative search. Firstly they had used KNN for classification if symptoms will showing more than one category then use differential diagnosis & LAMSTAR[3].

S. M. Kamruzzaman proposed system uses three steps for the classification of web pages. Firstly they had did automatically extraction of features by analyzing source code. In next step they had feeding values to neural network. And in last step they had define the category of web page according to specified eight domain[4].

Aijun An and Xiangji Huang had use HTML categorization method for classification of web pages. For classification of web page they had used ANN approach & the information in the form of HTML[5].

Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, Wei-Ying Ma had did classification of web links by using summarization technique. They had perform analysis of page layout with that they had extract main topics of web page using summarization algorithm which will enhance te performance of proposed system. proposed[6].

Arul Prakash Asirvatham Kranthi Kumar proposed system will classify the web pages by considering the structure of web document. They had also consider image characteristics with some broad categories [7].

Makoto Tsukada, Takashi Washio, Hiroshi Motoda proposed system did the co-occurrence analysis and by generates attribute for the classification of web pages. They had use machine learning technique to automatically classify the wepb page.[8]

Min-Yen Kan proposed a method which will classify the web pages based on URL. They had use two phase pipeline of word segmentation/ expantion and classification[9].

4. SYSTEM DESIGN

Objective

To develop a framework which enhance performance of web-page categorization to get efficient access to particular web-link with the help of artificial neural network.

4.1 Problem Statement

Create a system which can classify different websites or documents in specified domains by using Automatic features extraction through analyzing the home page source and all keywords must be processed by using Adaptive preprocessing. It should also be able to identify and avoid unrelated content on the page like advertisements and this classification should be done by using ANN.

4.2 Proposed System

The goal of proposed system is to classify the web links into its appropriate domain with the help of artificial neural network. The designed system will classify the web pages into five domain and that are Education, Entertainment, Business, Medical and Political. There are many technique available to classify the web links. To classify the web links it needs to be downloaded it cause to make extra burdon on server. It also require lots of time and unnecessary consumption of bandwidth. As system using ANN that means taking the benefits of unsupervised learning i.e. not to pre classify web pages at time of training to ANN as it require lots of efforts and may causes error in the classification. The system also uses lightweight crawling i.e. crawling operation allows the normalized operation of website [1].

We had designed our own crawler to extract the links which will fulfill all the three requirements specified above i.e. lightweight crawling, and classify without downloading. Our web crawler uses regular expression to extract keywords form web document. After extraction of keywords using instance processing concept keywords are made standardized and also defining the category of each word at same time by comparing it with predefined buzzword table contains keywords with their appropriate category. Then this keywords are stored into the database for learning purpose of ANN. Then using fixed windowing strategy for batch processing before feeding it to feed forward network.

Fixed Window Algorithm

Input : pre-processing window W_{pp}, predictor window W_{pr}, preprocessing model G,

Prediction model L, training data with (X, y), data for prediction(X_t, ……., X_{t+step})

Output: Predictions y'= (y'$_t$, ……., y'$_{t+step}$)

Fixed window strategies consist of three steps.

1. Train G_t with data (X, y)from interval[t-W_{pp}+1,t]

2. Train Lt with data(Gt(X),y) from interval[t-W_{pr}+1,t]
3. Predict y' = $L_t(G_t(X))$ for (X_t,……. ,X_{t+step})

After taking the values from batch processing ANN starts clustering. The ANN will calls back propagation encase of error or we can say difference in expected and actual value of output. If the ANN performs back propagation till maxrun over or the expected and actual result matches. Finally we will get the domain of entered link. That link with its domain identifier is stored into the database. If link for classification is of very first time then we get statistical result with pie chart indicating matching of keywords of specific domain according to which domain of that link is defined. If link provided for classification is already classified then output is shown from database.

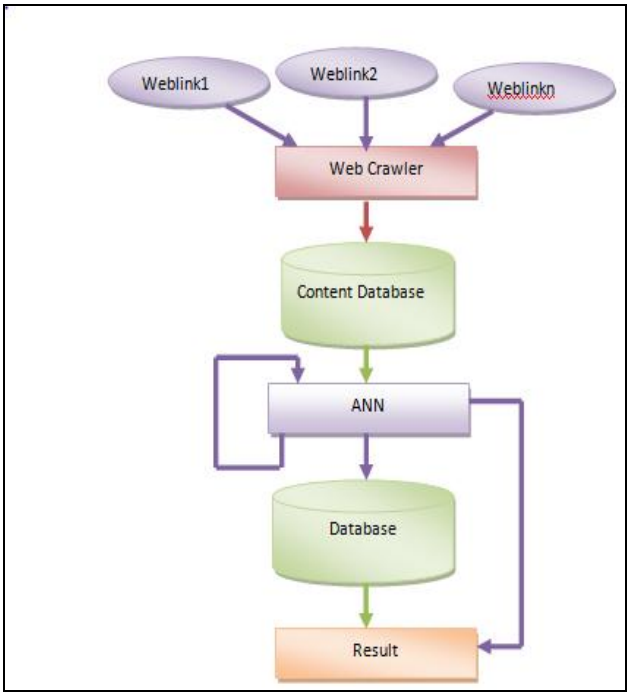

Fig1. System Architecture

Category	Words
Business	services , clients, vendors, investors, interest, loan, market policy, share market, property, privacy, infrastructure, business strategy,cost,stock,target etc.
Education	University, School, SSC, HSC, Engineering, Medical, Law etc.
Entertainment	Movies, Albums, Actor, Actress, TV Serial, Name of actors, movies and series can be included etc.
Medical	X-Ray, Neurology, Psychology, Pathology, Malaria, Dengue,Swine Flu etc.
Political	Election, Prime Minister, Mayor, MLA, Chief Minister, Name of politicians can be included etc.

Table1. Buzz Words

5. RESULT ANALYSIS

5.1 Execution of Keyword Extraction

System had web data extractor to retrieve various keywords from provided link document. These keywords are used by ADPP and ANN accordingly. To improve the performance of system extraction should be as fast as possible. The following graph indicates the number of keyword extraction with the provided link. There is five web link from which keywords are extracted by crawling provided link. The graph shows the variation in the extraction of keywords from different web sites. X-axis showing web sites to be handled by web crawler and Y-axis indicates number of keywords to be extracted by web crawler.

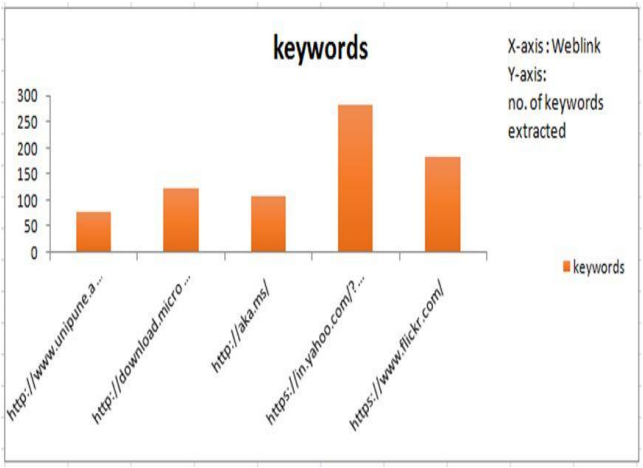

Fig2. Execution of keyword extraction

The following fig3. shows the variation of the execution time of different web sites. X-axis showing web sites to be handled by web crawler and Y-axis indicates time required to extract the keywords in milliseconds.

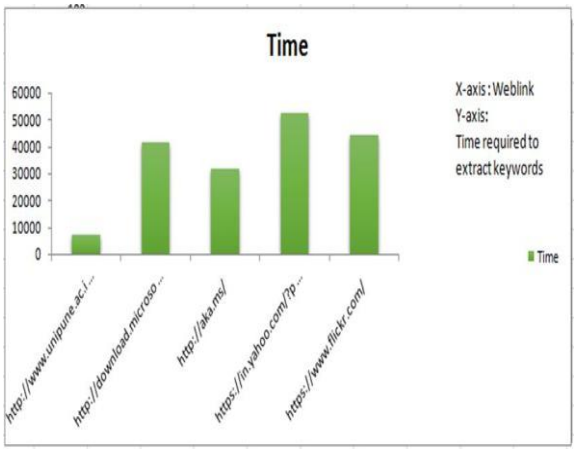

Fig3. Execution Time For Keyword Extraction

The following graph shows learning and execution Time of system. Learning time means the time which takes to start the system and actually started to extract keywords and the execution time means time which is calculated while domain of that link is classified. Five links are given and respective learning time and execution time is calculated. The fig4. shows weblinks on X-axis and Y-axis represents the values of learning time and execution time.

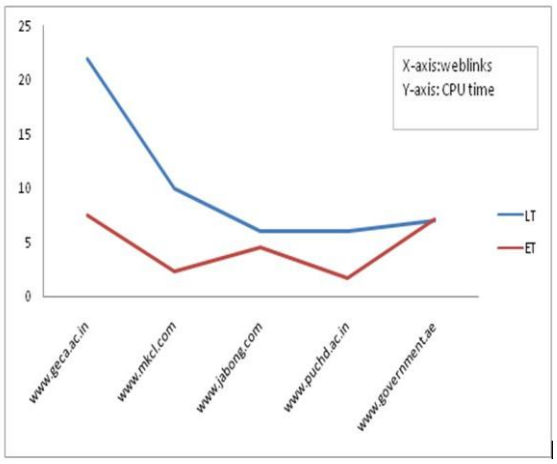

Fig4. Learning and Execution time for system

The accuracy of the system measures with two parameters and that are Mean Square Value and second is Pearson coefficient. The following graph shows the accuracy of the system with the same measures. In that we had did comparison of our system with VLBR algorithm & LM, BR, SCG algorithm with training testing and validation values.

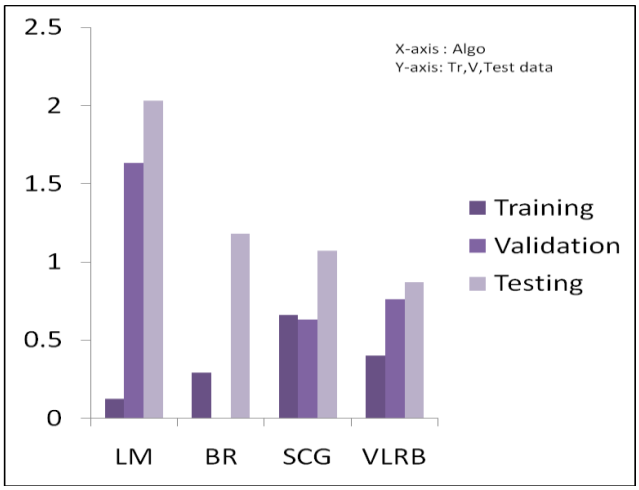

Fig5. Comparison for Mean Square error value.

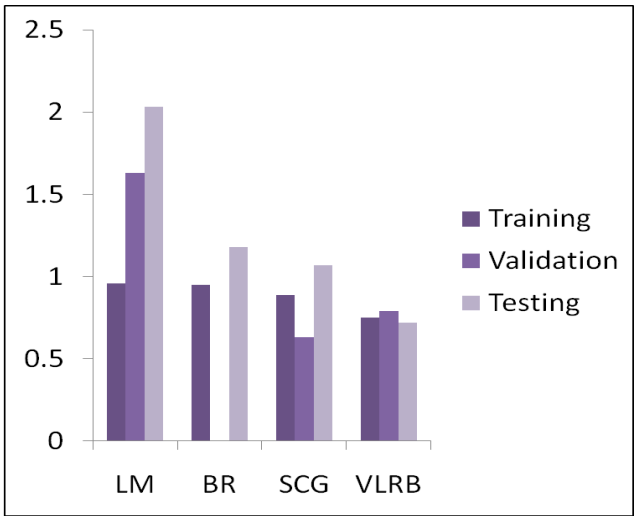

Fig6. Comparison for Pearson coefficient value.

6. CONCLUSION

The proposed system identification and classification of web pages with specified domain using ANN proposes a system for is providing enhanced way of categorization of web pages according to domain. In the proposed approach features are extracted by analyzing the source of a web page. By going through the tags of the source file the features can easily get. The site structure is defined by the internal or external links used in the page. By analyzing the reference or information pages (science/education/job site) we will get more external links than the commercial web sites (business and economy, news and media, sports sites). In the proposed approach we are defining some buzzwords that keep a site in to a certain class. As the frequency of buzzwords increases from same class; the probability of that site to be a part of that class also increases. The input values are decided after calculating the buzzwords to apply it to input layer of ANN. It should also be able to identify and avoid unrelated content on the page like advertisements and this classification should be done by using ANN and all keywords must be preprocessed to categorize the web page and create a system which can classify different websites or documents in specified domains. Existing approach for web page classification is studied and to enhance the performance of clustering the web pages into five domains by proposing new system with the help of ANN. The main objective is to provide an efficient way for categorization of web pages. Preprocessing the web pages will facilitate the different search engines to classify the web pages with more efficiency and also to provide a rich web directory. In future system can be used in DNS server application. System can replace the WWW with domain type.

7. ACKNOWLEDGMENTS

The authors would like to thank Shri Chairman Groups and Management and the Director/Principal Dr.Uttam Kalawane, Colleague of the Department of Computer Engineering and Colleagues of the varies Department the D.Y.Patil School of Engineering and Technology, Pune Dist. Pune Maharashtra, India, for their support, suggestions and encouragement.

8. REFERENCES

[1] Inma Hernndez, Carlos R. Rivero , David Ruiz , Rafael Corchuelo, "CALA: An unsupervised URL-based web page classification system", 2013 Elsevier B.V.http://dx.doi.org/10.1016/j.knosys.2013.12.019

[2] Indre Zliobait e and BogdanGabrys, Adaptive Preprocessing for Streaming Data,IEEEtransactions on knowledge and data engineering, vol. 26, no. 2, February 2014.

[3] Rahul Isola, Rebeck Carvalho, Amiya Kumar Tripathy,"Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR, and k-NN", IEEE transactions on information technology in biomedicine, vol. 16, no. 6, November 2012.

[4] S. M. Kamruzzaman (Jan 2006)'Web Page Categorization Using Artificial Neural Networks'Proceedings of the 4th International Conference on Electrical Engineering & 2nd Annual Paper Meet 26-28 January, 2006

[5] Aijun An and Xiangji Huang, "Feature selection with rough sets for web page categorization", York University, Toronto, Ontario, Canada.

[6] Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, Wei-Ying Ma, (Jully 2004)'

Web-page Classification through Summarization' SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK. Copyright 2004 ACM 1-58113-881-4/04/0007

[7] Arul Prakash Asirvatham Kranthi Kumar. Ravi,' Web Page Classification based on Document Structure' International Institute of Information Technology Hyderabad, INDIA 500019.

[8] Makoto Tsukada, Takashi Washio, Hiroshi Motoda,' Automatic Web-Page Classification by Using Machine Learning Methods' Institute of Scientific and Industrial Research, Osaka University Mihogaoka, Ibaraki, Osaka 567-0047, JAPAN.

[9] Min-Yen Kan, (May 2004)' Web page categorization without the web page' WWW2004, May 17–22, 2004, New York, New York, USA.ACM 1-58113-912-8/04/0005. Osaka University Mihogaoka, Ibaraki, Osaka 567-0047, JAPAN