Identification and Classification of Web Pages with Specified Domain

Poonam Nagale
Department of Computer Engineering, Dr.D.Y.Patil
School of Engineering & Technology,University of Pune,
India

Alka Vishwa
Professors, Department ofComputer Engineering,
Dr.D.Y.Patil School of Engineering & Technology,
University of Pune, India

ABSTRACT

Internet is very large source of information. But these flow of information need to be controlled in various organizations, i.e. in companies job portals and personal mail services are blocked, in colleges entertainment related websites are blocked. Consider college scenario, Admin have to keep watch on site the student accessing. He uses proxy services and firewall on sites that are not allowed to student access. But as per the growth of internet every day new sites launched in the market. It is not always feasible to admin to keep track on that, also every time we have to pay for proxy for each new site as well as it is somewhat time consuming. So, we are clustering the web links into five domain and the keywords must be preprocessed by Adaptive preprocessing technique to increase the performance of the system.

Keywords

Web mining, Feature extraction, ANN, Preprocessing data.

1. INTRODUCTION

We There is tremendous growth in internet users and the service providers to those users. It becomes necessary to identify and classify the web pages. WebPages are designed by human being only; to classify the web page manually will take lots of time it also effected by costing factors and it is quite boring job. So there is need of technique which will automatically classify the WebPages according to domain to reduce time and cost factors also too increase the accuracy of the work.[2][3] A most of research efforts had been taken which makes predictive model modify to ever growing changing environment. As the data grows over time, predictors need to be updated, else they become giving a less accurate output. The most of the predictors presume that the coming data must be in preprocessed form or it is built in part of learning algorithm. In real time application, the major step of data mining is that preprocessing the data. As number of people interacting with the system the data comes from may be redundant or less accurate or frequently noisy.

Data mining experts attest that data preparation requires 80-90 percent time of data mining project, so we can assume that modeling requires 10 percent. As data changing with time predictor adaption for the maintaining the better accuracy is not only the case but we have to follow the approach that will tie preprocessing with adaptive preprocessor. The main goal of adaptive system is to handle modified data. A preprocessing component of adaptive processing system with two main connection.

1. Preprocessor may need consistent feedback over time from predictor to decide upon adapting.
2. Preprocessor produces consistent mapping feature over time which modify the raw data, then it is used by predictor.

Adaptive preprocessing can be represented with meaningful scenario for that we have to cluster the adaptive learning approaches. These approach narrate technique supporting adaptive predictors, but they immediately convert for application to adaptive preprocessors. Adaptive learning consist of two models First, incremental learning model and Second, replacement model. Incremental learning model can include new data into existing model, while replacement learning approach will ditch the old model and learn new one from search on new data. Incremental learning model classify into instance level, batch level and ensemble level. In this paper we are considering only instance level and batch level. In instance level, updating the parameter of model is through the extracted information from one entering data point. In this approach we need one new data point to judge current accuracy. In batch level, models parameters are modified later when number of entering data points have been seen.[1] In this paper we are using ANN as predictor to predict the class of web links.ANN is a computational system inspired by the Structure, Processing Method, and Learning Ability of a biological brain. As Artificial Neural Network (ANN) is provides Massive Parallelism, Distributed representation, Learning ability, Generalization ability, Fault tolerance facility.

ANN also provide the function such as 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 [2][3][4][5][6][7].

2. LITERATURE SURVEY

All IndreZliobait[1] and BogdanGabrys[1] proposed mechanism which preprocess the data. This mechanism involves two methods one is instance processing [1][8][9], in this approach we need one new data point to judge current accuracy and second is batch processing [1][10][11] which are responsible for performing the operation on the preprocess data.

S. M. Kamruzzaman proposed [2] a technique of web page classification in three progressive stages, first they examine the source code for instinctive draw out the features. The second stage chooses the input values to the neural network. The third stage taking decision about a particular web page that is in which class it will drop out of eight predefined classes.

Aijun An [3] and Xiangji Huang [3] proposed mechanism of web page classification which uses HTML classification technique. In this with the help of HTML information seen in web page they are classifying the webpage with ANN classification technique.

Dou Shen [4], Zheng Chen [4], Qiang Yang [4], Hua-JunZeng [4], Benyu Zhang [4], Yuchang Lu [4], Wei-Ying Ma [4] proposed mechanism of summarization which is used to categorize the web page. With the help of Web summarization algorithm, they have carried out the analysis of page-layout for the draw out of main topic of web page to increase the classification accuracy.

Arul Prakash [5],Asirvatham [5], Kranthi Kumar[5] proposed a mechanism for categorizing web pagesautomatically consequently to structure of web document and it also consider the image characteristics for categorization into a few broad categories.

Makoto Tsukada [6], Takashi Washio [6], Hiroshi Motoda[6] proposed mechanism which uses is co-occurrence investigation to spawn an characteristic and also repeatedly classify webpage according to machine learning practice.

Min-Yen Kan [7] proposed mechanism to classify the web page into predefined category which will explore the utilization of URLs by a two-phase channel of word segmentation/expansion and categorization.

3. PROPOSED SYSTEM

To accomplish the requirement of clients several classification mechanisms are conversed by the diverse researchers.

Since there are tremendous increment in web pages every day, an proficient mechanism is projected here which will help to cluster the web pages supporting the five features dig out from a page and the classification is done in five successive stages. By examining about 500 web pages it is believed that all the web developers and the designers everlastingly try to represent the motto and the theme of the organization. The theme is expressed by giving the complete composition of the home page. The designer is always paying attention in to keep the tourist busy so they can spend more time in his site. So he designs the home page of any site with additional skill and assembles the structure of home page to accomplish the intent as well as the user approval [2][3]. So the five features are cached that will make the site unlike from another site. The features are structure of home page that will be represented with the ratio of internal and external links, number of dynamic/static pages to be used, image occurrences will found, animations availability is present in web page and the previously defined catchphrase. In this proposed approach, from different web directory five major classes are elected. The neural network are educated and experienced by using those classes [12][13]. Fig.1. shows the block diagram of system architecture. Fig.2. shows Adaptive learning technique. 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. For the network we can use a 5-3-5 architecture. The input vector of this network consists of 5 elements where each neuron represents one element. In this architecture one hidden layer with 3 neurons are used. Output of the network consists of 5 neurons, to show the output pattern based on our five classes. By five neurons we can classify the web pages into five categories and that are Education, Research, Entertainment, Medical, Political. There are lots of possibilities to be fraction value to come as output. To overcome this situation we are converting the values from 0.0 to 0.49 into 0 and the values from 0.50 to 1.0 into 1. Table1. Shows the list of some buzzwords. Following stages are used in the proposed approach:

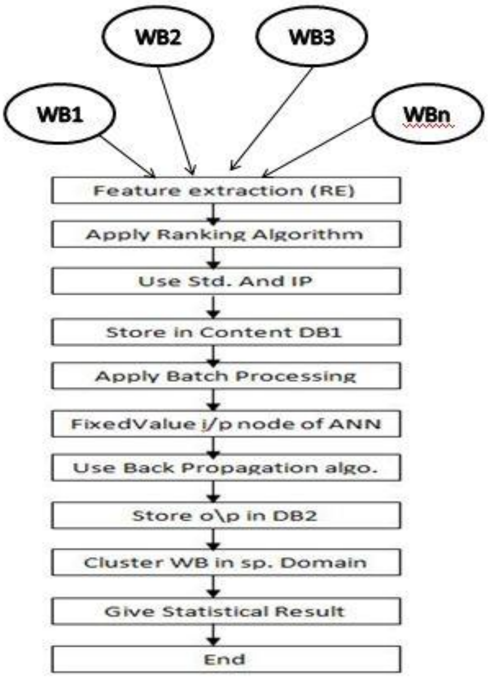
1. Analyzing home page source extract the features of it automatically.

2. Applying preprocessing and standardization methods i.e. take an example poonam and punam name should be considered as same and it will also see for the synonyms. We are using incremental learning because it includes new data into the old module. In instance processing need one new data point to judge current accuracy. Fig.3. shows hierarchy of adaptive learning.

3. At the input nodes of the networks unchanging the values.

4. Here also Applying standardization and another preprocessing technique which wait till the no. of data points are incoming and then model is modified Batch processing.

5. Categories web pages by the neural networks.



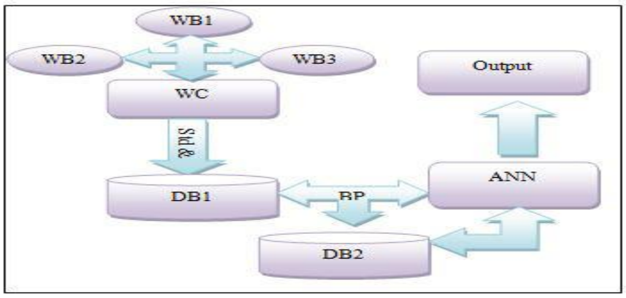


Fig. 1. System Architecture

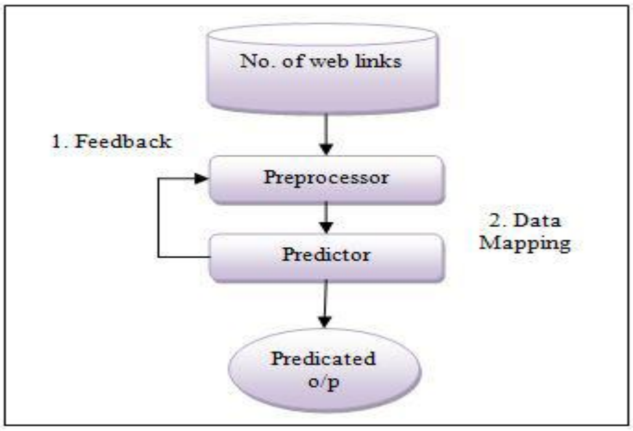


Fig. 2. Adaptive Learning Model

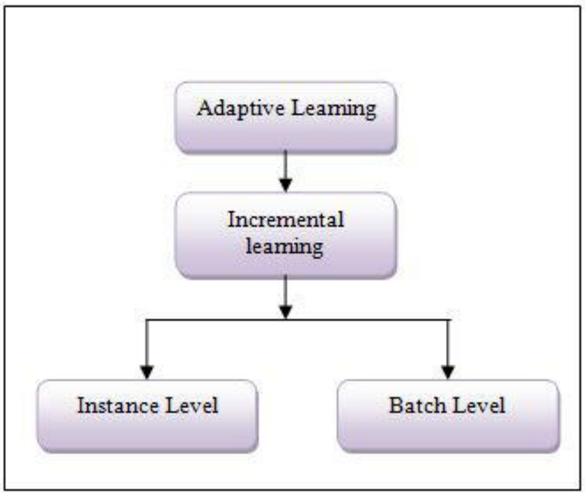
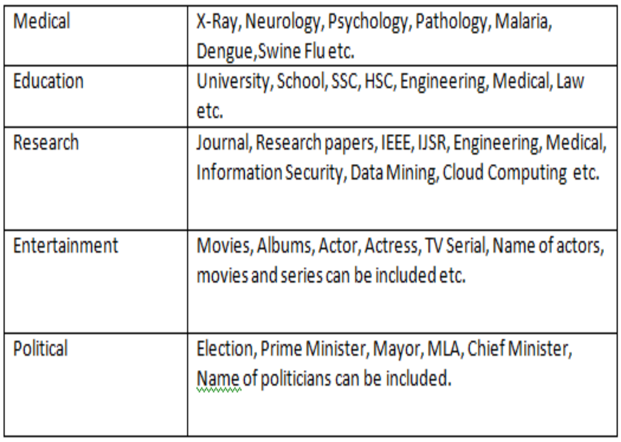


Fig.3. Hierarchy of Adaptive



Medical	X-Ray, Neurology, Psychology, Pathology, Malaria, Dengue,Swine Flu etc.
Education	University, School, SSC, HSC, Engineering, Medical, Law etc.
Research	Journal, Research papers, IEEE, IJSR, Engineering, Medical, Information Security, Data Mining, Cloud Computing etc.
Entertainment	Movies, Albums, Actor, Actress, TV Serial, Name of actors, movies and series can be included etc.
Political	Election, Prime Minister, Mayor, MLA, Chief Minister, Name of politicians can be included.

Table I. Buzzwords used in clustering

4. MATHEMATICAL MODEL

Place Let S = {WS,WC,U,N,D,L,A,}
where
1. WS is Set of weblinks of web source
WS = {L1,L2_ _ __Ln,}
where L is htpp link
2. WC is a set of web crawler
WC = {WC1,WC2_ _ __WCn, }
where WC is web crawler.
3. U is set of user
U ={U1,U2_ _ __Un,}
4. N is set of Neurons
N ={N1,N2_ _ __Nn, }
5. L is a set of Neuron Layer
L = {LN1,LN2_ _ __LNn,}
where LN is Neuron Layer.
6. D is set of database
D = {DK,}
where DK is keywords data.
7. A is admin which is unit set

5. PERFORMANCE ANALYSIS

Experiment was conducted with 300 web links to get the required result. This graphs satisfies the first feature extraction module. The input to web crawler is source web link also giving the buzzwords to extract the link having the same buzzword. And calculating number of links extracted and time required for extraction. In first graph the experiment is conducted with source web links and the number of links extracted from the provided source link. The X-axis indicates provided source link and Y-axis indicates the number of links extracted. The second graph shows relation between the source link and the time required to extract links from provided source links. The X-axis indicates the source link provided to web crawler and Y-axis indicates time(ms) required to extract those links.\

Source	Links	Time
http://www.unipune.ac.in/	78	7178
http://download.microsoft.com	124	41473
http://aka.ms/	108	31473
https://in.yahoo.com/?p=us	283	52351
https://www.flickr.com/	182	44498

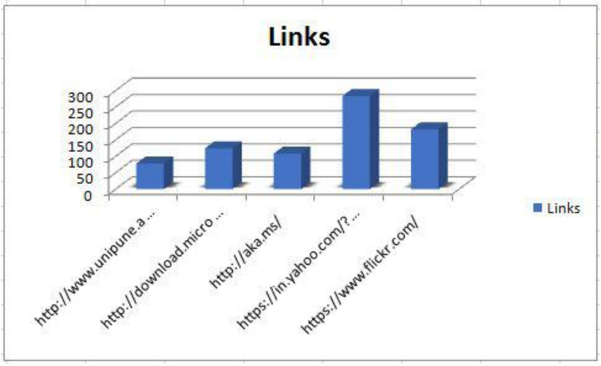


Fig. 4. no. of web link extraction

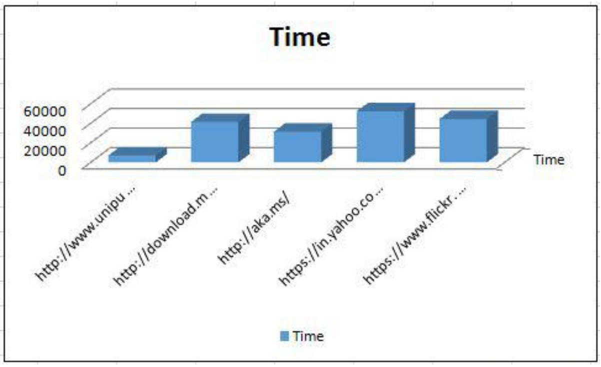
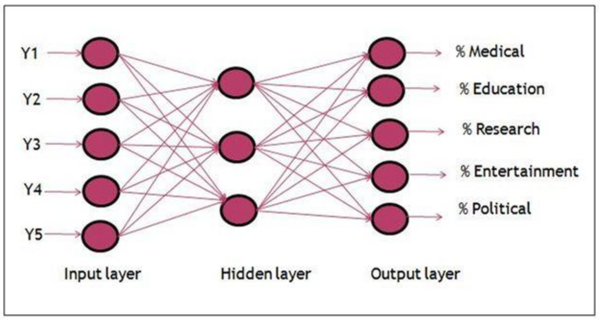


Fig. 5. Time(ms) to extract web links

6. CONCLUSION

Existing approach for web page classification is studied and to enhance the performance of clustering the web pages into five domain. We can use ANN to classify different URLs. Along with this we are going to use Adaptive preprocessing to decide the importance of any word during classification (Clustering) of URLs. Domain clustering can help search engines to improve their skills and also help firewalls to filter specific type of traffic. The proposed system is also capable of identifying and avoiding distinct contents like advertisement. Serve. In future system can be used in DNS server application. System can replace the WWW with domain type.



7. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We are thankful to the authorities of Savitribai Phule Pune University for their constant guidelines and support. Finally, we would like to extend a heartfelt gratitude to friends and family members. It provides for me massive joy to express my deepest feeling of appreciation and earnest because of my very regarded and regarded guide Prof. Alka Vishwa, Asst. Educator Asst. Professor Department of Computer Engineering, MAHARASTRA, INDIA for their profitable direction, support and help for finishing this work. Their valuable recommendations for this entire work and co-agent conduct are truly recognized. I likewise wish to express my appreciation to (Mrs. Roshni Raut and Dr. Uttam B. Kalwane) for his kind hearted backing and direction.

8. REFERENCES

[1] Indre Zliobait e and Bogdan Gabrys “Adaptive Preprocessing for Streaming Data”, IEEE transactions on knowledge and data engineering, vol.26, no. 2, february 2014

[2] S. M. Kamruzzaman “Web Page Categorization Using Artificial Neural Networks”,NetworksProceedings of the 4th International Conference on Electrical EngineeringJanuary, 2006.

[3] Aijun An and Xiangji Huang,“Feature selection with rough sets for webpage categorization”, York University, Toronto, Ontario, Canada. 2009.

[4] Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, Wei-Ying Ma, “Web-page Classification through Summarization”, SIGIR04, Sheffield, South Yorkshire, UK. Copyright 2004 ACM 1-58113-881-4/04/0007, July 2529, 2004.

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

[6] Makoto Tsukada, Takashi Washio, Hiroshi Motoda, “Automatic Web- Page Classification by Using Machine Learning Methods” [1] Institute of Scientific and Industrial Research, Osaka University Mihogaoka, Ibaraki,Osaka 567-0047, JAPAN.

[7] Min-Yen Kan,“Web page categorization without the web page” WWW2004, New York, New York, USA.ACM 1-58113-912-8/04/0005. Osaka University Mihogaoka, Ibaraki, Osaka 567-0047, JAPAN, May 1722, 2004.

[8] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda ,“New Ensemble Methods for Evolving Data Streams” , Proc. 15th ACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining(KDD 09),pp. 139-148, 2009.

[9] E. Ikonomovska, J. Gama, and S. Dzeroski,“Learning Model Trees from Evolving Data Streams”, Data Mining Knowledge Discovery,vol. 23, no. 1, pp. 128-168, 2011..

[10] P. Kadlec and B. Gabrys,“Architecture for Development of Adaptive on-Line Prediction Models”, Memetic Computing,vol. 1, no. 4, pp. 241- 269, 2009.

[11] M. Masud, J. Gao, L. Khan, J. Han, and B. Thuraisingham,“ Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints”, IEEE Trans. Knowledge and Data Eng.,vol. 23, no. 6, pp. 859-874, June 2011.

[12] D. Boley, M. Gini, R. Gross, E-H. S. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore,“Partitioning-based clustering for web document categorization”, Decision Support System.,1999.

[13] SPrabhakar Gold wasser and Eli Uphal Chandra Chekuri Michale,Stamford University,“Web Search Using Automatic Categorization”, IBM alamden Research Center, 650 Harry Road,San Jose CA.s