ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

A Survey on Domain Clustering Technique with Adaptive Preprocessing

Shraddha Kharche[1], A. P. Kankale[2]

P.G. Student, Department of Computer Science Engineering, RSCOE, buldhana, Maharastra, India[1]

Professor, Department of Computer Science Engineering, RSCOE, buldhana, Maharastra, India[2]

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. In this paper we use ANN i.e. Artificial neural network and also applying adaptive preprocessing technique to preprocess the data before if feedin to the ANN to enhance the accuracy features from web links.

KEYWORDS: Web mining, Feature extraction, ANN, Preprocessing data.

I. INTRODUCTION

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] .

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 [3][4][5][6]. It is ok to categories the web pages manually. But it is time consuming and too tired some job to search a page of intended class.

So we have designed a technique which will categories web pages using a technique for web page categorization using artificial neural network(ANN) through automatic feature extraction and also using a instance and batch processing which are preprocessing technique. In this paper we are using ANN as predictor to predict the class of web links. 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. Consider a college scenario, they want to provide access only that will relate to education domain and there is ban on usage of network sites that are may be media or prone sites. it is not always possible to keep track of sites that will accessed by students, by an admin.

As there will be new site launched by the companies for the same per day. So, we will identify and classify that site in a domain to block that site by service provider. So for the better use of resources we have proposed a new and efficient technique that will facilitate the user for categorization the web pages. This work covering five major classes of web pages which may include business & economy, education, government, entertainment, job search.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

II. LITERATURE SURVEY

IndreZliobait[1] and BogdanGabrys[1] proposed mechanism which preprocess the data. This mechanism involves two methods one is instance processing [1], in this approach we need one new data point to judge current accuracy and second is batch processing [1]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 choose 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 classifyizng the webpage with ANN classification technique.

Dou Shen [4], Zheng Chen [4], Qiang Yang [4], Hua-Jun Zeng [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 thedraw 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 pages automatically 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.

III. IMPLEMENTATION

To accomplished the requirement of user several categorization techniques are discussed by the different researchers. Because of tremendous increment in the web pages every day, an efficient technique is proposed here which will classify the web pages based on the five features extracted from a page and the categorization is done in five successive stages. By analyzing about 500 web pages it is found that all the web developers and the designers always try to enlist the motto and the theme of the organization. The theme is expressed by giving the total structure of the home page. The designer is always interested in to keep the visitor busy more time in his site. So he tries to design the home page with extra skill and build the home page structure to fulfill the intention as well as the user satisfaction.

So the five features are catching up that make the site different from other types of site. The features are home page structure, which is the ratio of internal and external links are present, amount of dynamic/static pages are used, frequency of images will found, availability of animations is present in web page and the predefined buzzwords. In this proposed approach, five major classes are selected from different web directory. The buzzwords used in the clustering are medical, education, research, entertainment, and political.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

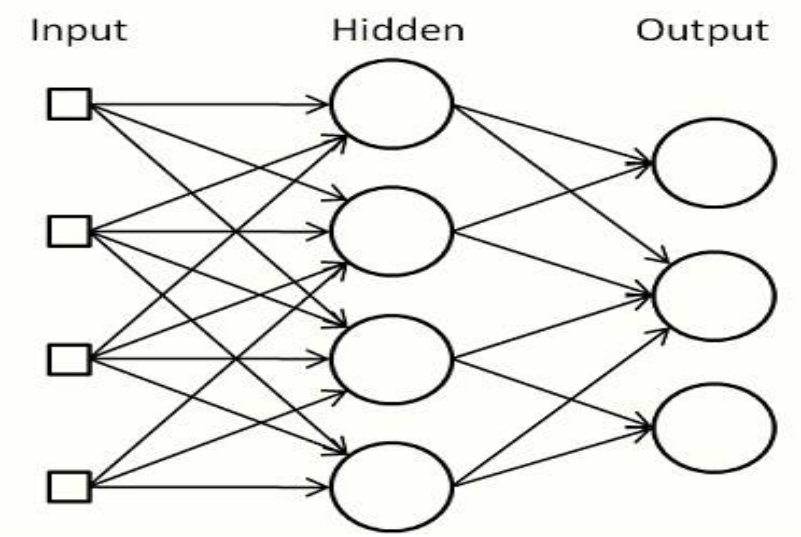

Figure 1: Artificial Neural Network (ANN)

The neural network are trained and tested by using those classes. The proposed approach is done through the Following stages:

1. By analyzing home page source automatically Extract the features.
2. Applying standardization and Instance processing.
3. At the input nodes of the networks fixed the values.
4. Applying standardization and Batch processing.
5. Categories web pages by the neural networks.

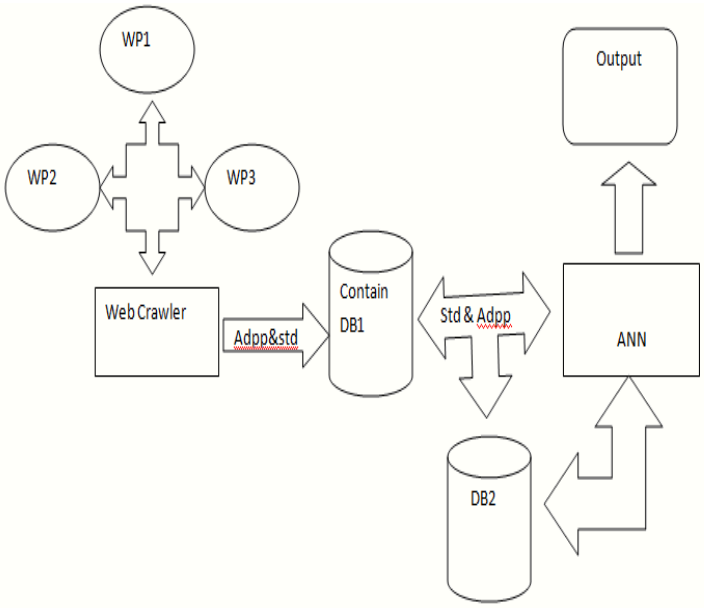

Figure 2: System architecture

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

IV. ADVANTAGES

1. To block the certain domain web pages.
2. We categories the web page and block it by service provider to avoid the misuse of resources.
3. Use this application for the specific searching.
4. This system gives age based response.

V. CONCLUSION

The existing approaches to categorized the web page is studdied and Create a system which can classify different websites or documents in specified domains by using Automatic features extraction through analyzing the home page source. 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 processed by using Adaptive preprocessing.

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 Engineering January, 2006.
[3] Aijun An and Xiangji Huang,"Feature selection with rough sets for web page 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. Thuraising-ham,"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.