International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Review of Domain Clustering Using Adaptive Preprocessing

Poonam Nagale[1], Alka Vishwa[2]

[1]Department of Computer Engineering, Dr. D. Y. Patil School of Engineering & Technology, University of Pune, India

[2]Professors, Department of Computer Engineering, Dr. D. Y .Patil School of Engineering & Technology, University of Pune, India

Abstract: ***A huge amount of data has available on the WWW. It is necessary to categories that web pages which will be a challenging job for efficient use and easy to access to web page. In previous paper the Author uses ANN to classify the web page into eight domain categories. In this paper we are using an ANN and also applying the Preprocessing technique to preprocess the data before it feeding to the ANN to enhance accuracy by extracting features from web links. The web pages are generated dynamically in response to users queries through Web based search categorized into the five domains.***

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

1. Introduction

The continuous growth in accessing World Wide Web Make it essential to invent new technique that will be able to automatically categories the web pages into certain classes or categories. This will be guideline not only for the end users and server but also for the search engines to get the required site with less time requirement and improved accuracy. We have studied Automatic categorization of web pages, and realize that most of these categorization techniques are usually based on similarity between documents contents or their structures [1] [2].

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 proposed a technique which wills 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. 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 launching 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.

2. Literature Survey

Indre_ Zliobait e and Bogdan Gabrys proposed[1] technique to preprocess the data which involves two methods first is instance processing and second is batch processing to enhance the operation performed on the preprocess data.

S. M. Kamruzzaman proposed [2] a system of web page categorization in three progressive stages; in first stage they analyze the source code for automatically extracting the features. The second stage fixes the input values to the neural network. The third stage will decides the class of a particular web page out of eight predefined classes.

Aijun An and Xiangji Huang proposed [3] web page classification by using HTML categorization method .In this they are using HTML information present in web page to classify the webpage with ANN approach. Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, Wei-Ying Ma proposed[4] summarization technique to classify the web page. They have used Web summarization algorithm, they have did the analysis of page-layout for the extraction of main topic of web page to enhance the accuracy of classification.

Arul Prakash Asirvatham Kranthi Kumar proposed[5] a method to automatically classify of web pages according to structure of web document and the characteristics of images into a few broad categories.

Makoto Tsukada, Takashi Washio, Hiroshi Motoda proposed [6] technique which is done by co-occurrence analysis to generate an attribute and also automatically classify webpage according to machine learning technique. Min-Yen Kan proposed [7] technique for web page categorization which will explores the use of URLs by a two-phase pipeline of word segmentation/expansion and classification.

3. Proposed System

To fulfill the demand of user several categorization techniques are discussed by the different researchers. Because of continuous growth of 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 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.

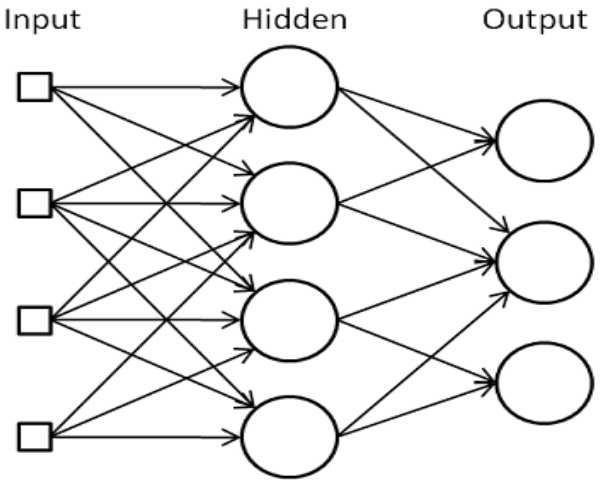

Figure 1: Artificial Neural Network

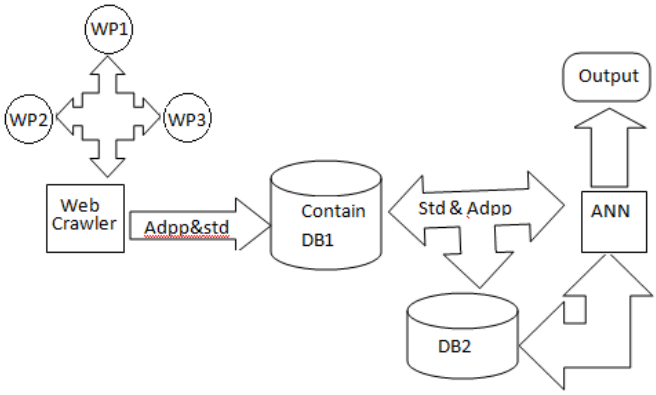

Figure 2: Basic Architecture of Proposed System

4. Applications

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.

5. Conclusion

In this way we have studded the existing approaches to categorize the web page 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'Proceedings of the 4th International Conference on Electrical Engineering & 2nd Annual Paper Meet 26-28 January, 2006
[3] Aijun An and Xiangji Huang, "Feature selection with rough sets for web page categorization", York University, Toronto, Ontario, Canada.
[4] Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, Wei-Ying Ma, ' Web-page Classification through Summarization' SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK. Copyright 2004 ACM 1-58113-881-4/04/0007
[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' 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, May 17–22, 2004, New York, New York, USA.ACM 1-58113-912-8/04/0005. Osaka University Mihogaoka, Ibaraki, Osaka 567-0047, JAPAN.

Author Profile

Poonam Suresh Nagale Research Scholar Dr. D.Y.Patil School of Engineering & Technology, Pune, University of Savitribai phule pune. He received B.E. in Computer Engineering from Parvatabai Genba Moze college of Engineering, Wagholi, Pune. From Savitribai phule pune University. Currently she is pursuing M.E. in computer engineering from Dr.D.Y.Patil School of Engineering & Technology, Pune, University of Savitribai phule pune.

Alka Vishwa (1986) was born in Ajmer, India in 1986.She received her degree of B.Tech from Rajasthan University in 2008 and M. Tech from Gyan Vihar University, Jaipur in 2013. She has published over 8 refereed journal and conference papers in the areas of neural networks. Some of her representative published papers list is as follows: "Classification of arrhythmic ECG data using machine learning Techniques" published in IJCSITS, "Speckle noise reduction in ultrasound images using wavelet thresholding" published in IJARCSSE, "Modified method of prediagnosis of lung cancer using forward ANN" published in IJCSE, "Modified method of speckle noise reduction in ultrasound images" published in IJISA. Her research interests include artificial neural networks and nanotechnology.