A Proposed Technique for Finding Pattern from Web Usage Data

Karunendra Verma
Sir Padampat Singhania University
Udaipur, India
Email:k.verma2006@gmail.com

Dr. Prateek Srivastava
Sir Padampat Singhania University
Udaipur,India
Email:prateek.srivastava@spsu.ac.in

Dr.Prasun Chakrabarti
Sir Padampat Singhania University
Udaipur, India
Email:prasun.chakrabarti@spsu.ac.in

***Abstract*—** There are various ways of web page classification but they take higher time to compute with lesser accuracy. Hence, there is a need to invent an efficient algorithm in order to reduce time and increase web page classification result. Artificial Immune System (AIS) which has the characteristic of high self-adaptation and self-construction inspired from the function of biological immune system. An ensemble approach of AIS and tree based classifier has used the hybrid approach. This inspired the scholars to use such hybrid approach for Structure based web page classification.

Keywords- *Artificial Immune System, Web Structure Mining, Classification.*

*****

I. Introduction

Now a day's World Wide Web has become very popular and interactive for transferring information. The web is huge, diverse and active thus increases the scalability, multimedia data and temporal matters. The growth of the web has its outcome in a huge amount of information that is now freely offered for user's access. Several different kinds of data have to be handled and organized in a manner that they can be accessed by the users effectively and efficiently.

The web is a collection of interrelated files on one or more web servers. Web mining is the application of data mining techniques to extract knowledge from web data including web documents, hyperlinks between documents, usage logs of web sites etc.

Web mining is broadly divided into three categories: web structure mining, web content mining and web usage mining.

AIS contain the ability of learning, recognizing, memorizing and characteristic extraction. At present, the research findings of the artificial immune algorithm based on danger model are almost talking about realization of algorithm, and application in Fraud detection. The central idea in the Danger Theory is that the immune system does not respond to non-self but to danger. It fundamentally supports the need for discrimination, instead of responding to foreignness, the immune system reacts to danger signal.

II. Review of Literature

Kovacevic et al. [9] proposed a new hierarchical representation that includes browser screen which coordinates with every HTML object in a page. Using visual information one is able to define heuristics for the recognition of common page areas such as header, left and right menu, footer and centre of a page. In the initial experiments the author shows that using heuristics defined objects are recognized properly in 73% of cases. Finally, show that a Naive Bayes classifier, taking into account the proposed representation. Zou et al. [16] have found that due to the presence of the noisy data there is a need for classification of the web page for real world applications. A method which will properly ensure the classification is the support vector machine because it has the capability of generalization. Author's suggested method provides a way which will increase the accuracy of classification by combining the support vector machine concept with the K-nearest neighbour techniques. Tomar et al. [4] introduce the concept of a classification tool for web pages called Web Classify, which uses modified naive Bayesian algorithm with a multinomial model to classify pages into various categories. In this research experimental result along with the classification accuracy analysis with increasing vocabulary size, was also shown. Ryan et al. [10] studied the area of genre categorization has an emphasis on retrieving the features such as text from the specific documents. Since the main aim of work is taken into account whether visual properties of HTML web page can significantly improve the categorization of pulverous genres. Apparently, it seems that it would place a major challenge and will be also beneficial to retrieve those visual characteristics which catching the layout feature of genres. The bulk of web pages produced from various business websites and manually categorized into genres. The three different characteristics are compared side by side a). With the textual characteristics b). With the HTML characteristics c). Visual characteristics. Author's work has the ability to prove that by using HTML characteristics and URL characteristics helps in increasing the accuracy of classification as compared to textual alone. Thus, it also seems that by adding the visual characteristics, it increases the pulverous classification. Kang et al. [7] present a analysis on mining web data from the numerous available data on WWW. As the web pages are not fully structured so it becomes difficult to determine from the informative block methods which provide the useful data extraction from the useless data such as advertisements which is more important. In this proposed method author introduce a web page classification in form of blocks by constructing a tree alignment model that indicate the HTML feature and a vector model that represents a feature of blocks. Thus, by constructing the single classifier it becomes difficult to classify a block accurately. To overcome this problem in proposed method author uses the multiple classifiers one for each training data set and classification method succeeds by combining all of them. Mun et al. [11] found that the size of web page increases a lot as the number of offered services as well as link increases and then due to their accessing speed decreases. The author uses the link graph arrangements for troubleshoot this problem. By introducing this link graph system author enables to reduce the load of server to a greater extent. Rathod [13] shows outlines of three different modes of web mining, namely web content mining, web structure mining and web usage mining. The development and application of Web mining techniques in the context of web content usage and structure data will lead to tangible improvements in many web applications from search engines and web agents to web analytics and personalization. Gowri et al. [3] described a brief survey about the existing approach in web services composition. The main research areas in web services are related to discovery, security, and composition. Among all these areas, web services composition turns out to be a challenging one because within the service-oriented computing domain, Web service composition is an effective recognition to satisfy the hastily changing requirements of business. Therefore, the Web service composition has unfolded itself broadly in the research side. However, the current attempts to classify Web service composition are not appropriate to the objectives. This paper proposes a novel classification matrix for Web service composition, which distinguishes between the context and technology dimensions. The context dimension is aimed at analyzing the QoS influence on the effort of Web service composition, while the technology dimension focuses on the technique influence on the effort. Finally, this paper provides a suggestion to improve the quality of service selection which participates in the composition process with C skyline approach using agents. Sarac et al. [14] worked on the firefly algorithm (FA) inspired by the flashing behavior of fireflies, which belongs to the category of Meta heuristic algorithm. It flashes primary intention to attract other fireflies through a signaling system Jain et al. [2] proposed a new method "Intelligent Search Method (ISM)". In this method author proposes to index the web pages using an intelligent search strategy. This new method integrated with any of the page ranking algorithms to produce better and relevant search results. Keller et al. [8] present a GRABEX method for extracting navigational block types based on the link patterns. The method was applied to mine breadcrumb navigations. Analyzing to which other navigational block types the GRABEX method can be applied is also interesting for future work. An author believes that paginations or previous/next navigations can be mined as well if suitable graph generation methods are implemented. The GRABEX method can also be extended to mine non-navigational page elements if graphs are not generated from hyperlinks but from other structures e.g. text or linked images. Jose et al. [6] show the Rough set theory applications in various domains like business, medicine, commerce, telecommunication and many other fields. The results of this approach can be used for target advertising because advertisers can post their advertisements on content pages especially pages in lower approximation. This also helps to identify the most preferred content by a user because users spent more time on potential pages. Ye et al. [15] improved and proposed a brand new method of semantic relevancy algorithm based on the Wikipedia hyperlink network, integrating the semantic information in the paging network and the category network reasonably to carry out semantic correlation calculation. The method fully taps the rich semantic information in Wikipedia and regards semantic correlation as an adjustable parameter in the algorithm. So it can be suitable for the flexible use in different of fields, being more consistent with human's analysis and cognitive habits. After comparative analysis of other methods such as algorithm based on the Wikipedia page network only, based on the Wikipedia category network only and simple weighted algorithm by summing of the results obtained by the two networks, author proved its reasonableness.

He et al. [5] work based on the fact that the web is a collection of various web documents. The classification of a web document is meant for three things mainly: indexing, search and retrieval. There is a difference between web classification and text classification. This difference is due to the structure of the web documents. These differences could be one or more of the following: meta data, the title of the document and various links available in the document etc. In this paper, authors have chosen either of the following methods such as information gain (IG), DF-thresholding (DF), $\chi 2$ -test (CHI) or term strength (TS) criterion, Information gain and $\chi 2$ -test for feature selection for classification. After feature selection, this paper uses Support Vector Machine (SVM) classifier for classification. The method confirms, quantifies and extends previous research by introducing a new structure-based method for description and classification of web documents. Compared to traditional web document classification methods, combining the full text with structure information gains nearly 6% accuracy improvement in the case of similar categories and 3.7% accuracy improvement in the case of distinct categories.

III. RESEARCH GAP

The literature review entails that; the classification done on the dataset of web structure is optimized by structure based web document analysis. However, beside these described techniques there are various other ways also to perform web structure based classification. Certainly, web structure based classification gives better result in association with feature selection results because it finds various features in the record of dataset. But while using this technique with simple web structure based classification, there is a scope of improvement in the following two concerns.

- Web structure based classification itself takes longer time to compute.
- The result of simple web structure based classification is not that much optimized.

And the reasons behind these two concerns are the accuracy of the classification method. It is proposed to improvise these concerns to get better efficiency in this work.

IV. METHODOLOGY

To overcome above limitations, I divided my work into following activities:

Activity I:

- To study various web structure mining algorithms.
- To design a Java code for extracting structure base feature selection from database of web pages. (Figure 1)

Activity II:

To design and apply artificial immune system concept on the dataset to find irrelevant data. On the basis of structure based feature selection it gives a desire dataset. (Figure 1)

➢ *Artificial Immune System Algorithm:*

Step 1: Preparation of a dataset from above steps (activity 1) which will be the outcome of the structure analysis.
Step 2: Extracting the desired information based on the structure of the web pages.
Step 3: To identifying the pages based on structure information which cannot provide the desired information.
Step 4: Removing these pages from the dataset to make better outcome.

Activity III:

- Designing and implementing structure finding algorithm such as information gain (IG), DF-thresholding (DF), χ2 -test (CHI), term strength (TS) criterion or our own methods. (Figure 1)
- Appling tree based classifier using WEKA tool for finding various classes/group of web pages based on Structure of the web pages.

➢ *Tree Based Classifier Algorithm:*

Step 1: Reading dataset.
Step 2: Dividing the dataset into two parts: Training and Testing dataset.
Step 3: Training the classifier with training dataset.
Step 4: Preparing the rules and model for the tree based classifier.
Step 5: Applying the prepared model on testing Dataset.

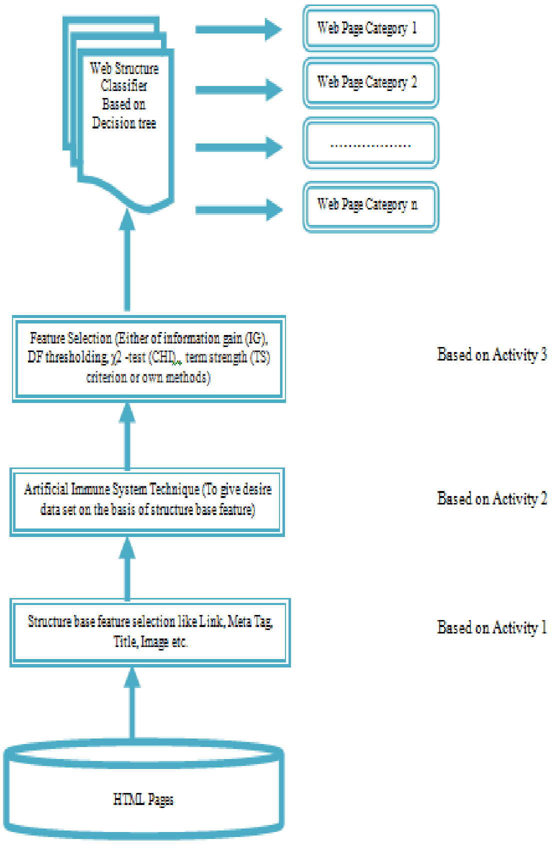
Web Structure Classifier Based on Decision tree
Web Page Category 1
Web Page Category 2
....
Web Page Category n
Feature Selection (Either of information gain (IG), DF thresholding, χ2 -test (CHI),, term strength (TS) criterion or own methods)
Based on Activity 3
Artificial Immune System Technique (To give desire data set on the basis of structure base feature)
Based on Activity 2
Structure base feature selection like Link, Meta Tag, Title, Image etc.
Based on Activity 1
HTML Pages

Figure 1 Overall Proposed Architecture

V. CONCLUSION AND FUTURE WORK

This paper presents a structure-based method using Artificial Immune System to achieve high accurate web document classification. With the help of structure information, which includes META tags, TITLE, descriptions of links and alternative texts of images, approach will evaluate using the dataset, and will demonstrate the advantages of structure-based classification for both similar categories and distinct categories. We can also compare to traditional web document classification method, and compare the results.

ACKNOWLEDGMENT

This research is the part of PhD thesis and I am thankful to our guide Dr. Prateek Srivastava, Professor in department of computer science & engineering, Sir Padampat Singhania University , Udaipur (Rajasthan) to encourage and guide in this topic which will helpful to speedup structure based web page classification for fast search.

REFERENCES

[1] A. Sadegh, H. Abolhassani, R. Hossein and N Behroo, “Web page classification using social tags,” IEEE International Conference on Computational Science and Engineering Volume 4 , 588-593 , Aug. 2009.

[2] A. Jain, R. Sharma , G. Dixit and V. Tomar, "Page ranking algorithms in web mining, limitations of existing methods and a new method for indexing web pages," International Conference on Communication Systems and Network Technologies, IEEE pp.640-645, Dec. 2013.

[3] Gowri, R. And Lavanya, R.,"A novel classification of web service composition and optimization approach using skyline algorithm integrated with agents," IEEE Computational Intelligence and Computing Research (ICCIC), 26-28 , Dec. 2013.

[4] G. S. Tomar, S. Verma and A. Jha,"Web page classification using modified naïve bayesian approach, "IEEE TENCON 2006 Hong Kong, pp.14-17 , Nov. 2006.

[5] H. Kejing and C. henyang ,”Structure-based classification of web documentsusing support vector machine,” IEEE, Proceedings of CCIS2016 215-219, 2016.

[6] J. Jeeva and P. SojanLal , "A rough set approach to identify content and navigational pages at a website, " IEEE , 2013.

[7] K. Jinbeom and C. Joongmin,"Block classification of a web page by using a combination of multiple classifiers," IEEE Networked Computing and Advanced Information Management Volume 2, pp.290-295 , Sept.2008.

[8] K. A. Matthias and H. Hannes,"GRABEX: A graph-based method for web site block classification and its application on mining breadcrumb trails, "WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT) ,IEEE, pp.290-297, 2013.

[9] K. Milos , D. Michelangelo , G. Marco and M. Veljko ,"Recognition of common areas in a web page using visual information: a possible application in a page classification ,"IEEE Data Mining ,pp. 250-257, 2002.

[10] L. Ryan , C. Michal , and Y. Lei ,"Using visual features for fine-grained genre classification of web pages, "IEEE Hawaii International Conference on System Sciences, Proceedings of the 41st Annual 1-10 , 7-10 Jan. 2008.

[11] M. Yilhyeong,L. Minkyung and C. Dongsub,"Classification of web link information and implementation of dynamic web page using Link Map System ," IEEE Granular Computing 26-28 , Aug. 2008.

[12] Q. QIAN ,J. LI , J. CAI , R. ZHANG and M. XIN ," An anomaly intrusion detection method based on pagerank algorithm," International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, IEEE 2226-2230 , 2013.

[13] R. Dushyant,"A review on web mining," International Journal of Engineering Research and Technology (IJERT) , 2012.

[14] S. Esra, O. Selma Ayse ,“Web page classification using firefly optimization, ” Innovations in Intelligent Systems and Applications (INISTA), IEEE International Symposium, 2013.

[15] Y. Feiyue, Z. Feng ,L. Xiangfeng and X. Lingyu," Research on measuring semantic correlation based on the wikipedia hyperlink network," IEEE, pp. 309-314, 2013.

[16] Z. Jia-qi , C. Guo-long and G. Wen-zhong,"Chinese web page classification using no se-tolerant up port vector machines, "Natural Language Processing and Knowledge Engineering, IEEE NLP-KE, pp. 785-790 , 30 Oct.-1 Nov. 2005.