International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 10 3018 – 3021

Artificial Immune System based Firefly Approach for Web Page Classification

Miss. Rupali A. Mulay
Department of Computer Science & Engineering,
NIIST, RGPV University,
Bhopal, INDIA
rupali.mulay@gmail.com

Astt. Prof. Abhishek Singh Chauhan
Department of Computer Science & Engineering,
NIIST, RGPV University,
Bhopal, *INDIA*
abhichauhan78@gmail.com

Abstract—WWW is now a famous medium by which people all around the world can spread and gather the information of all kinds. But web pages of various sites that are generated dynamically contain undesired information also. This information is called noisy or irrelevant content. Web publishing techniques create numerous information sources published as HTML pages. Navigation panels, Table of content, advertisements, copyright statements, service catalogs, privacy policies etc. on web pages are considered as relevant and irrelevant content. This paper discusses various methods for web pages classification and a new approach for content extraction based on firefly feature extraction method with danger theory for web pages classification.

Keywords—*Web Page; Classification; Firefly; Danger Theory; Feature Selection,; Artificial Immune System.*

*****

I. INTRODUCTION

Over the past decade, web users have witnessed an exponential growth in the number of web pages accessible through popular search engines. Organizing the large volume of web information in a well-ordered and accurate way is critical for using it as an information resource. One way of accomplishing this in a meaningful way requires web page classification. Web page classification addresses the problem of assigning predefined categories to the web pages by means of supervised learning. This inductive learning process automatically builds a model over a set of previously classified web pages. The learned model is then used to classify new web pages.

With the popularity of the Internet, we entered the era of data explosion. While networking of thing, cloud computing, big data processing technology have been developed considerably, but World Wide Web technology is still the most widely used form of data in people's daily working and life. As the explosive growth of Web information, when network information is retrieved, the problems we faced is not whether the information can be accessed, but how exactly to get the desired Web page, it is usually from a large number of search results document, to choose needed, valuable information [1].

Traditional search tools can not working effectively, because they usually return a large number of search results [2]. Although we have entered some carefully chosen keywords, but the search engines usually return a huge number of web pages as a search result, for example, a common situation is to return dozens or even hundreds of URL hyperlinks. Only through manual operation, we can get the really need Web page with manual browsing filter.

In contrast to the traditional benchmark datasets, web directories generally have complex statistical properties. This makes large-scale hierarchical web page classification significantly different from traditional text classification and from web page classification with limited categories and documents. Web directories usually exhibit a spindle distribution having more categories and documents in the middle of the hierarchy than at either the upper or the lower levels of the hierarchy.

This paper attempts to apply artificial ammine system to improve the classification performance of Web information. We propose a effective feature selection method firefly in association with artificial ammine system for web classifier, establish a set of practices model to implement Web search results for the automatic classification of information, hoping to bridge the differences between existing search technologies and the needs of users [3].

The article is organized as follows: Section 2 is the literature review. The dataset used for this research is discussed in Section 3. The experimental setup used for this research is discussed in Section 4. Section 5 is the results and discussions. The recommendations of this research are summarized in Section 6.

II. ARTIFICIAL IMMINE SYSTEM

The problems found in a self and non-self are quite similar to those encountered in a Biological Immune System (BIS), since both of them have to maintain stability in a changing environment. Due to numerous desirable characteristics of the natural immune system, such as diversity, self tolerance, immune memory, distributed computation, self-organization, self-learning, self-adaptation, and robustness, BIS has attracted many researchers' attention [4] [5]. At the same time, Artificial Immune System (AIS) have become an increasingly popular computational intelligence paradigm [6][7].

Artificial Immune System (AIS) are still relatively young and the natural immune system (NIS) is one of the most complex systems under active study by biologists, there are some distinct viewpoints about the main goal of the NIS. These ideas and understandings are extremely important for AIS researchers and designers.

The main two distinct viewpoints are between self, non-self theory and danger theory. The classical immunology stipulates that an immune response is triggered when the body encounters something non-self or foreign [8]. This viewpoint is generally accepted by immunologists, and the models are created by AIS researchers based on this approach. A lot of question marks arise from this viewpoint, and a new theory called Danger Theory has been developed. The main idea behind danger theory is that the immune system does not respond to non-self but to danger. Similarly like the self non-self theories, it fundamentally

 ISSN: 2321-8169
3018 – 3021

supports the need for discrimination. However, it differs in the answer to what should be responded to. Instead of responding to foreignness, the immune system reacts to danger [9].

And researchers can use the concept of AIS in whole including self and non-self for outsider in web page classification.

III. Web Data Mining And Web Classification

Data mining is the study of data-driven techniques to discover patterns in large volumes of raw data. Web mining can be referred as the transformation of the data mining techniques to web data. Web mining has three distinct phases involved – content, structure and usage mining of web data. Mining the content involves extracting the relevant information, structure mining studies the structure and prototype and usage mining is the analysis of the discovered patterns. Web Usage Mining (WUM) is all about identifying user browsing patterns over WWW, with the aid of knowledge acquired from web logs. The outcomes of the WUM can be used in web personalization, improving the performance of the system, modification of the site, business intelligence, usage characterization etc.

The working of WUM has three steps – preprocessing of the data, pattern discovery and analysis of the patterns. Results of the pattern discovery directly influenced the quality of the data processing. Good data sources not only discover quality patterns but also improve the WUM algorithm. Hence, data preprocessing is an important activity for the complete web usage mining processes and vital in deciding the quality of patterns. In data preprocessing, the collection of various types of data differs not only on type of data available but also the data source site, the data source size and the way it is being implemented.

IV. Firefly Feature Selection

It is a meta heuristic algorithm, inspired by the flashing behaviour of fireflies [12]. To attract other fireflies is the main aim of firefly's flash. Xin-She Yang formulated this firefly algorithm by assuming: 1.All fireflies are unisex, so that one firefly will be attracted to all other fireflies; 2. Brightness make them attractive accordingly, and for any two fireflies, the less brighter one will attract (and thus move) to the brighter one; here, distance increases makes decreases of brightness; 3.If there are no fireflies brighter than a given firefly, it will move randomly. So objective function must have brightness component. Recent studies show that FA is particularly suitable for nonlinear multimodal problems.

V. Architecture Of AIS-Firefly Feature Selection Based Classification

All methods of abnormal behaviour detection involve the gathering and analysis of information from various areas within a computer or network to identify possible threats posed by hackers and crackers inside or outside the organization. In this area, IDSs fall into two categories according to the detection approaches they employ, namely i) anomaly detection and ii) Misuse detection. Misuse detection identifies intrusions by matching observed data with pre-defined descriptions of intrusive behaviour. Therefore, well-known intrusions can be detected efficiently with a very low false alarm rate. But this approach will fail easily when facing unknown intrusions.

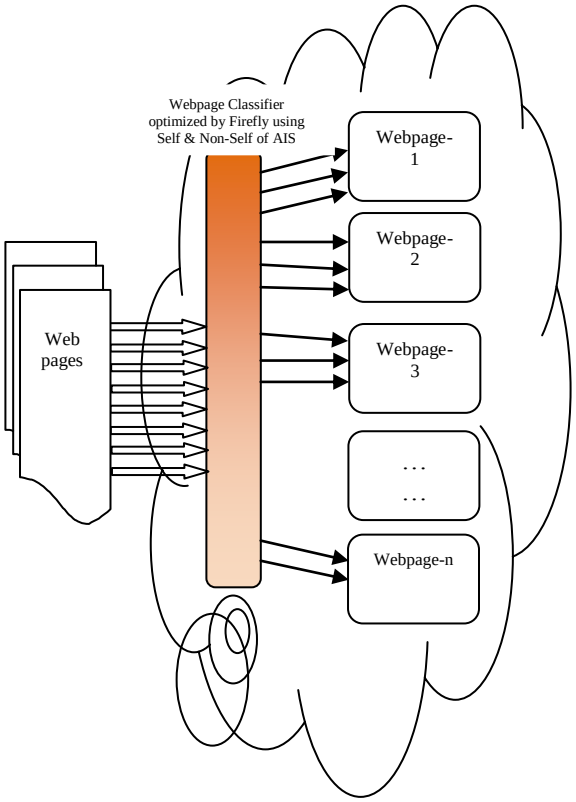

Figure 1: Architecture of classifier based on firefly feature selection motivated by Self and Non-Self concept of Artificial Immune System.

Natural or say our immune system work for protecting our self from various outsider unwanted bodies like viruses and bacteria[13]. System is made up of various things like abilities, including pattern recognition, inherent distributed parallel processing etc. There are variety of techniques available which could be useful to simulate the immune system that are aimed at resolve the concern at very large extend. Common techniques in this field that explain behaviour of the immune system are Clonal Selection, Negative Selection, Immune Network and self and Non-Self algorithms.

VI. Simulation Result

In this section, we will evaluate the performance of artificial immune system's self- non-self based firefly feature selection based classifier with simulation test.

A. Collecting Sample Information

For our artificial immune system's self- non-self based firefly feature selection based classifier can work correctly, researchers have to provide it with the learning sample and testing sample. It is the same with Biological Immune System, after training stage, the artificial immune system's self- non-

self based firefly feature selection based classifier can also get the proper results.

The generic search engine can be used to gather various subjects Web page documents, and from these documents the leaning sample and testing sample data can be obtained. The collecting process is as follows: first the keywords of subject are inputted by human or by system program, then the search engine often returns large numbers of Web pages. We often choose front fifteen pages as max relevant results.

To generate researcher observations, researchers have used the following setting: at first, through automatic method the ten categorical keywords can be retrieved from a Web document. Second, These ten categories are divided into further 30 sub categories. In each category there is sub categories of three. Researchers estimate subject class of the Web document carefully. The estimation is a simplex manual work. This procedure is repeated until we have a sufficient number of learning sample in the database. In our study, we have collected 334 Web pages.

B. Simulation Test

For every collected Web page, the 3 subject keywords must be extracted using firefly feature selection method. These keywords should reflect the main attributes of Web document Our Web Page classification system adopts statistical method to implement indexing process, which is based on calculating frequency of appropriate words as the keywords. If the document has keywords already, system adopt it's directly and select another keywords, until attaining 9 ordered keywords. Then researchers can code these keywords as a lighten the page for input variables of firefly, the Hash coding scheme can be chosen as one mapping or transforming method from character string to code number.

The system used for execution of the AIS based Firefly feature selection method with existing [11] methods is as follows:

TABLE 1: PREQUISITE SOFTWARE AND HARDWARE

Model	Pentium Dual Core CPU 2.20Ghz
RAM	4GB
32 Bit Operating System	
Windows 7 Ultimate	

In order to evaluate the classifier of Web Page classifier system, researchers use 334 Web pages for learning and 334 Web pages for testing. These 334 Web pages are collected by `Cornell University, Ithaca, NY`: [10]. To generate our test Web pages, researchers have used the following steps: at first, through statistical method the 30 keywords can be retrieved from a Web pages. Second, researchers estimate class of the Web pages carefully. The estimation is a simplex automatic work. In this study, researchers have collected 334 Web pages for learning and testing need.

Researchers compare the Artificial Immune System based Feature selection by firefly method for classification result with the web page classification by firefly optimization technique[11] result, and find that there are 334 Web pages rightly classified into correct subject class by proposed methods of feature selection in classifier. The test result shows that the average system performance of Artificial Immune System based Firefly Feature Selection optimization for classifier is in Table 2 and 3.

TABLE 2: EXECUTION TIME COMPARISON BETWEEN PROPOSED WORK WITH EXISTING WORK [11].

S. No.	Existing Method [11]	AIS based FS for Classier
1	10.579	7.541112
2	10.405	7.478047
3	10.326	7.36577
4	10.3393	7.396642
5	10.3931	7.387252

TABLE 3: CLASSIFICATION RESULTS COMPARISON BETWEEN PROPOSED AND EXISTING WORK[11].

S. No.	Existing Method [11]	AIS based FS for Classier
1	95.209581	95.726496
2	95.209581	95.726496
3	95.209581	95.726496
4	95.209581	95.726496
5	95.209581	95.726496

VII. CONCLUSION

After analysis of Table 1 and 2, which is output of the proposed Artificial Immune System based Firefly Feature Selection method for Classification is father better as it has two main advantages over existing method [11]. Researchers can show this betterment by two points, which are as follows:

1. On the basis of Classification Results
2. On the basis of Execution time of the methods.

So finally researchers have land up with a method which provides higher classification percentage with lesser execution time.

REFERENCES

[1] B. Liu, C. W. Chin and H. T. Ng, "Mining Topic-Specific Concepts and Definitions on the Web," in Proceedings of the twelfth international conference on World Wide Web

conference (WWW-2003), Budapest, HUNGARY, pp. 20-24, May 2003.

[2] U. Sukakanya and K. Porkaew, "A Framework for Automatic Classification of e-Business Web Content," in Proceedings of the Fourth International Conference on eBusiness, Bangkok, Thailand, pp. 19-20, November 2005.

[3] X. Liu, Y. Gong, W. Xu, and S. Zhu, "Document clustering with cluster refinement and model selection capabilities," in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, pp. 191-198, 2002.

[4] F. Sun, and F. Xu, "Antibody concentration based method for network security situation awareness," Proc. of the 3nd International Conference on Bioinformatics and Biomedical Engineering(iCBBE 2009), IEEE Press, pp. 1-4, June 2009.

[5] F. Sun, S. Cheng, "A gene technology inspired paradigm for user authentication," Proc. of the 3nd International Conference on Bioinformatics and Biomedical Engineering(iCBBE 2009), IEEE Press, pp. 1-3, June 2009.

[6] W. Zhang, C. Wu, and X. Liu, "Construction and enumeration of Boolean functions with maximum algebraic immunity," Science In China, Series F: Information Science, vol. 52, pp.32-40, January 2009.

[7] F. Sun, and Z. Wu, "A new risk assessment model for e-Government network security based on antibody concentration," Proc. of the 2009 International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government. pp. 119-121, December 2009.

[8] Perelson, A. S. & Weisbuch, G. (1997). "Immunology for Physicists", Rev. of Modem Physics, 69(4), pp. 1219-1267.

[9] P. Matzinger, (2002), "The Danger Model: A Renewed Sense of Self', Science, 296, pp. 301-305.

[10] Computer Science Department, Cornell University, Ithaca, NY 14853-7501, Fall 1996, http://www.cs.cornell.edu.

[11] Esra Saraç and Selma Ayşe Özel, "Web Page Classification Using Firefly Optimization," 2013 IEEE.

[12] Yang X. S., "Firefly algorithms for multimodal optimization", Stochastic Algorithms: Foundations and Applications, SAGA 2009. Lecture Notes in Computer Sciences, Vol.5792, pp.169–178.

[13] L. N. de Castro and J Timmis, Artificial Immune Systems: A New Computational Intelligence Approach, 2002.

AUTHORS

Rupali A. Mulay appered for M.Tech in Computer Science & Engg from Ragiv Gandhi Proudyogiki Vishwavidyalay University & received the B.E. degrees in Inforamtion Teechnology from Shri Sant Gadge Maharaj University in 2011.Presenting 2nd International Conference on ICIE 2013, , Review on Optimized Webpage Classification, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 4, July-August 2014 ISSN (2278-6856).

Abhishek Singh Chauhan (Asst. Prof.) appeared PhD. [Computer Science &

Engineering] from Bhagwant University, Ajmer (Rajasthan) & received M.Tech from Rajeev Gandhi Prodyogiki Vishwavidhyalaya, Bhopal (2012), MCA from IGNOU (2005), M.Com from Pt. Ravishankar Shukla University, Raipur(2000), PGDCA from Makhanlal Chaturvedi University, Bhopal (2001), GNIIT fromNIIT (Graduate from National Institute of Information Technology) (1999). A Comprehensive Survey on Frequent Pattern Mining from Web Logs. (IJAEA Jan 2011).A Novel Approach for web usage mining using Growing Neural Gas. (National Conference on Recent Trends in Mathematics & Computing (RTMC) April 2011).Detecting and Searching System for Event on Internet Blog Data Using Cluster Mining Algorithm

(Springer)http://www.springerlink.com/content/j13834wp348211j0/(Jan2012). Rule Based Lexical Dictionary based polarity analysis for reviews, International Journal of Emerging Trends in Electronics & Computer Science (Jun 2013).Secure Content Sniffing For Web Browsers : A Survey, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013.Review on Classification of Web

Log Data using CART Algorithm, International Journal of Computer Applications (0975 – 8887) Volume 80 – No 17, October 2013.Mining Association Rules from Infrequent Itemsets: A Survey, International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), Vol. 2, Issue 10, October 2013, (5801-5808), , Review on Optimized Webpage Classification, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 4, July-August 2014 ISSN (2278-6856).