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