EIN-WUM an AIS-based Algorithm for Web Usage Mining

Adel T. Rahmani and B. Hoda Helmi
Soft Computing Lab., Dept. of Computer Engineering, Iran University of Science and Technology
Narmak, Tehran 1684613114, Iran
rahmani@iust.ac.ir, bh_ helmi@comp.iust.ac.ir

ABSTRACT

With the ever expanding Web and the information published on it, effective tools for managing such data and presenting information to users based on their needs are becoming necessary. In this paper, we propose a new algorithm named "EIN-WUM" for Web usage mining based on artificial immune system metaphor. This algorithm introduces several novelties such as using danger theory, directed mutation and an enhanced immune network model. Experimental results show that The EIN-WUM algorithm can properly learn the frequent trends in noisy, sparse and huge Web usage data in single pass.

Categories and Subject Descriptors: I.5.2 [Pattern Recognition]: Design Methodology _ *Pattern Analysis*

General Terms: Algorithms, Design, Experimentation.

Keywords: Artificial immune system, Web usage mining, Directed mutation, Danger theory.

1. INTRODUCTION

Web Usage Mining is one of the applications of data mining to discover interesting usage patterns from Web. Artificial Immune System is a new, biologically inspired, paradigm for learning. There are a number of motivations for using the immune system as inspiration for Web usage mining which include recognition, diversity, memory, self regulation, noise tolerance and learning [1]. Artificial immune network models are developed by Timmis [2] and De Castro and Von Zuben [3] which have been used as biologically motivated approaches to data mining. But all the existent immune network models suffer from the high computational costs of calculating the interaction between antibodies. In EIN-WUM an Enhanced Immune Network model which is capable of learning the frequent patterns in single pass of data is presented to overcome the computational cost of existent immune network models. The rest of paper is organized as follows: in section 3 a brief pseudo code for EIN-WUM is presented. Next section danger theory, EIN and directed mutation are explained briefly and section 4 shows the experimental results.

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-130-9/08/07.

2. AN OVERVIEW ON EIN-WUM

A brief pseudo code of EIN-WUM is shown below:

1- Initialize antibodies using some of input data (sessions)
2- Construct neighborhoods using a simple clustring method.
3- Set the neighborhood threshold to average dissimilarity between cluster prototypes.
4- For each antigen (incoming sessions)
  4-1 Calculate danger level of antigen (interestingness of session), if danger level of antigen is more than a threshold continue, else goto 4.
  4-2 Present antigens to cluster prototypes.
  4-3 Choose the most activated neighborhood.
  4-4 If affinity between antigen and selected neighborhood prototype is less than neighborhood threshold add a new neighborhood with a copy of antigen to the network, update neighborhood information and goto step 4.
  4-5 Else calculate stimulation level of antibodies in the selectd neighborhood, update neighborhood information and update antibodies' vector according to flaw or supplementary state.
  4-6 Clone antibodies.
  4-7 After processing every T antigen, mutate antibodies, add the new antibodies to the network.
  4-8 Delete excess antibodies with least stimulation level, move stagnated antibodies to second memory.

3. ENHANCED IMMUNE NETWORK (EIN), DIRECTED MUTATION AND DANGER THEORY

The immune network theory was proposed as a way to explain the memory and learning capabilities exhibited by the immune system. In this paper we propose and use a new immune network model named EIN to fulfill our needs.
The general equation for calculating the stimulation level of antibodies in immune network models is composed of 3 terms. These terms are effect of j antigens on i^{th} antibody and excitatory and inhibitatory effect of neighboring antibodies on the i^{th} antibody. To model dynamics of the network and avoid the undesirable computation overload in existent models, we use neighborhood information. The neighborhood information are updated whenever a change occur

in the antibodies of the neighborhood and this update occurs incrementally. The neighborhood information used for calculating the dynamic of the network are similarity between antibodies in the neighborhood and the affinity between i^{th} antibody and the prototype of the neighborhood $pt(nbh_{Ab_i})$, so the stimulation of an antibody in the EIN is calculated through (1).

$$s_{Ab_i} = \frac{W_i + w_{i,j}}{\sigma_i^2} + \frac{w_{i,pt(nbh_{Ab_i})}}{\sigma_i^2} - similarity(nbh_{Ab_i}) \quad (1)$$

Danger theory (DT) introduced by Matzinger [4] attempts to explain the nature and workings of an observed immune response in a way different to the more traditional view. In the context of data mining, danger can be interpreted as interesting [5], when an antigen enters the EIN-WUM, if the antigen interestingness is more than a pre-specified threshold, danger signal is released and antibodies try to combat (learn) the antigen, otherwise the not-interesting antigen is discarded. This form of DT is implemented by definition of interest metric for each antigen [6]. Another form of DT is used to control the metadynamics of the immune network. This form of DT is implemented via directed mutation that will be described in next paragraph.

In this paper directed mutation is introduced to overcome the problems of blind mutation and to control the metadynamic of the network. This mechanism is designed as follow: As it is obvious matching between antigen and antibodies are not always complete. When URL_i exists in antigen and doesn't exist in antibody, a flaw in antibody is detected, so a -1 is added to the i^{th} element of antibody vector indicating the lack of URL_i in antibody. And when URL_i doesn't exist in antigen but exists in antibody, a supernumerary state is detected and a $+1$ is added to the i^{th} element of antibody vector indicating the surplus of URL_i in antibody. After every T antigens are processed a measure for each of antibodies is calculated based on number of times that either flaw or supernumerary states has occurred. For each element in antibody vector, if number of flaw or supernumerary state occurred is more than a pre-specified threshold (strain bound), that element is called strained bit. Based on number of strained bits in antibody, a new metric named Degree of inept of an antibody is calculated as (2):

$$IneptDegree(Ab) = \frac{strainedBit(Ab)}{L} \quad (2)$$

If inept degree of an antibody is more than a threshold, then a copy of the antibody is created and the copy undergoes mutation. The mutation operation is a simple point mutation on strained bits of the antibody. This mechanism is a new from of using danger theory to control the metadynamics of the network. Removing the excess and useless antibodies is done by defining an upper bound on number of antibodies in the network.

4. EXPERIMENTAL RESULTS

We used the evaluation method developed in [7] to show the ability of EIN-WUM for summarizing the huge, noisy and sparse data such as Web usage data. The log of 10 day user access to the Website of music machine on the server www.hyperreal.org is used. The log consists of 220146 request and 19542 sessions. During these 10 days, 4500 URLs had been accessed. A single pass over the entire data sets

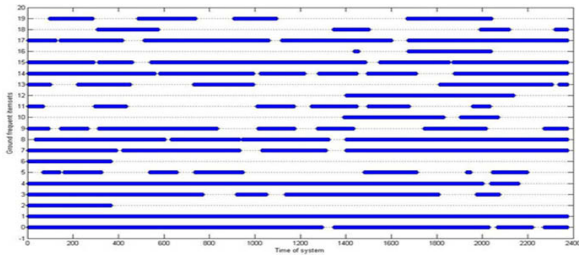

Figure 1: Accurate and complete antibodies relative to the ground truth profiles

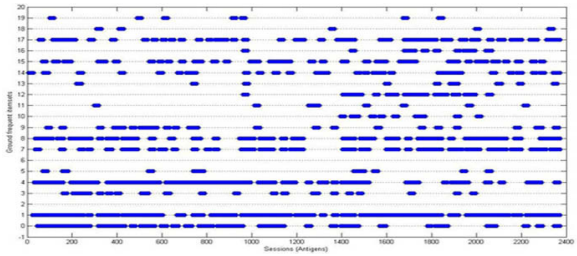

Figure 2: Distribution of input data

(with non-optimized C++ code) took 5 min on a 2 GHz Pentium 4 PC. In evaluating the goodness of the learned antibodies, it must be shown that the final antibodies (figure 1) should represent the input data stream (figure 2) with respect to its ground truth profiles as accurately as possible and as completely as possible [7].

5. CONCLUSIONS

In this paper a new robust algorithm with several novelties such as incorporating danger theory, new immune network model and directed mutation is presented. This algorithm is able to learn frequent patterns of Web usage data in single pass of input data. The main factor of the algorithm that has made it capable of learning the frequent patterns in single pass is its rich and manageable immune network.

6. REFERENCES

[1] A. Secker. *Artificial Immune Systems for Web Content Mining: Focusing on the Discovery of Interesting Information.* University of Kent in Canterbury, UK, 2006.

[2] J. Timmis. *Artificial Immune Systems: A Novel Data Analysis Technique Inspired by The Immune Network Theory.* University of Wales, Wales, 2000.

[3] L. N. de Castro and F. J. Von Zuben. *Artificial immune system: a new computational intelligence approach.* Springer-Berlin, 2002.

[4] P. Matzinger. The Danger Model: A Renewed Sense of Self. *Science*, 296:301–305, 2002.

[5] A. Secker and A. A. Freitas and J. Timmis. A Danger Theory Inspired Approach to Web Mining. In *Proc. 1st Int. Conf. on Artificial Immune Systems (ICARIS), Lecture Notes in Computer Science 2787*, pages 156–167. Springer-Verlag, 2003.

[6] B. H. Helmi and A. T. Rahmani. An AIS Algorithm for Web Usage Mining with Directed Mutation. In *Proceedings of the World Congress on Computational Intelligence (WCCI'08)* (To be published). 2008.

[7] O. Nasraoui and C. Rojas and C. Cardona. A Framework for mining evolving trends in Web data stream using dynamic learning and retrospective validation. *Computer Networks*, 50:1488–1512, 2006.