The Fuzzy Artificial Immune System: Motivations, Basic Concepts, and Application to Clustering and Web Profiling

Olfa Nasaroui*, Fabio Gonzalez, and Dipankar Dasgupta
Dept. of Electrical and Computer Engineering* and Intelligent Security Systems Research Lab,
Division of Computer Science
The University of Memphis
e-mail: {onasraou, fgonzalz, ddasgupt}@memphis.edu

Abstract - The human immune system can be seen as a complex network structure that is able to respond to an almost unlimited multitude of foreign invaders such as viruses and bacteria. Hence, this parallel and distributed adaptive system promises tremendous potential in many intelligent computing applications, including Web mining. Some of these immunity-based techniques involve the development and analysis of algorithms that can identify patterns in observed data in order to make predictions about unseen data. In this paper, we introduce several new enhancements to deal with some of the weaknesses of previous artificial immune system models. In particular, we address the uncertainty and fuzziness inherent in the matching process that takes place between antibodies and antigens. This problem is handled by introducing a fuzzy artificial immune system. A fuzzy artificial immune system mimicking the body's adaptive learning and defense mechanism in the face of invading biological agents is used as a monitoring and learning system for a Web site in the face of all incoming Web requests.

1. INTRODUCTION

Most living organisms exhibit extremely sophisticated learning and processing abilities that allow them to survive and proliferate generation after generation in their dynamic and competitive environments. For this reason, nature has always served as inspiration for several scientific and technological developments. One such natural system is the natural immune system that can be seen as a parallel and distributed adaptive system [2] that has tremendous potential in many intelligent computing applications. This is because the immune system exhibits the following points of strength: recognition, feature extraction, diversity, learning, memory, distributed detection, and self-regulation [2,3]. The immune system uses combinatorics to construct pattern detectors efficiently. Moreover, the detection/recognition process is highly distributed in nature. Based on these underlying mechanisms, an intelligent computational technique has been developed for pattern recognition and data analysis [11].
One of the data repositories, affecting every aspect of our life lately, is the World Wide Web. In addition to its ever-expanding size and lack of structure, the WWW has not been responsive to user preferences and interests. One way to deal with this problem is through personalization. *Mining* information from the user's interaction is another approach towards personalization. Perkowitz and Etzioni [16] proposed adapting Web pages based on a user's traversal pattern. In [1], associations and sequential patterns between web transactions are discovered. Most of the above efforts have relied on relatively simple techniques which can be inadequate for real user profile data since they are not resilient to the "noise" typically found in user traversal patterns. To handle possibly unknown noise contamination rates in Web data, Nasraoui et al. [13] have proposed mining the Web log data using a fuzzy relational clustering algorithm based on a robust estimator. In this work, they have also proposed the formal definition of a "robust" user profile and "robust" quantitative evaluation measures. To deal with the fuzzy nature of Web data and to automatically determine the number of clusters, profiles were extracted [14,15] using an unsupervised fuzzy relational clustering algorithm based on competitive agglomeration.
The rest of the paper is organized as follows. In Section 2, we present an overview of the natural immune system. In Section 3, we review some of the current artificial immune system models. In Section 4, we present our fuzzy AINE model. In Section 5, we illustrate using the fuzzy AINE model for clustering. In Section 6, we describe our artificial immune system inspired approach to Web usage mining. In Section 7, we illustrate the performance of our approach in extracting session profiles from the access log file of a real Web site. Finally, in Section 8, we present our conclusions and future prospects.

2. THE NATURAL IMMUNE SYSTEM

The natural immune system is a distributed novel-pattern detection system with several functional components positioned in strategic locations throughout the body. The main purpose of the immune system is to recognize all cells (or molecules) within the body and categorize those cells as *self* or *non-self*. The non-self cells are further categorized in order to stimulate an appropriate type of defensive mechanism. The immune system learns through evolution to distinguish between foreign antigens (e.g., bacteria, viruses, etc.) and the body's own cells or molecules.
The *lymphocyte* is the main type of immune cell participating in the immune response that possesses the attributes of specificity, diversity, memory, and adaptivity. There are two subclasses of the lymphocyte -- T and B, each having its own function. In particular, *B-Cells* secrete antibodies that can bind to specific antigens.

3. ARTIFICIAL IMMUNE SYSTEM MODELS

Artificial Immune Systems emerged in the 1990s as a new computational research area. Artificial Immune Systems link several emerging computational fields inspired by biological

0-7803-7280-8/02/$10.00 ©2002 IEEE

behavior such as Artificial Neural Networks and Artificial Life. This new immuno-informatics approach has been investigated and practical applications developed notably by Dasgupta and Forrest [2,3,4,5,7,8]. One can view that the immune system (lymphocyte elements) can behave as an alternative biological model of intelligent machines, in contrast to the conventional model of the neural system (neuron elements). The Artificial Immune System has enabled the use of computational models of information processing in immunological inter-actions, with practical applications to many problems, e.g. pattern recognition, data mining, computer security, and fault detection [2].

In their attempt to apply immune system metaphors to machine learning, Hunt and Cooke [9] based their model on Jerne's *Immune Network theory* [12]. The system consisted of a network of B cells used to create antibody strings that can be used for DNA classification.

Timmis et al's resource limited AIS (AINE) model [17] consists of *ARBs (Artificial Recognition Balls)*, each consisting of several *identical* B cells, a set of antigen training data, links between ARBs, and cloning operations as usual. Each ARB represents a *single n*-dimensional data item that could be matched by Euclidean distance to an antigen or to another ARB in the network. A link was created if the affinity (distance) between 2 ARBs was below a Network Affinity Threshold threshold parameter, NAT, defined as the average distance between all data items in the training set. Each member of the antigen training set is matched against each ARB based on Euclidean distance. This affects the ARB *stimulation* level which is inversely related to its average distance from the antigen set. When its stimulation level exceeded a certain threshold, the ARB was *cloned* and mutated. Unlike the previous models, the ARBs compete for a *finite* number of resources (B cells). This is done by initially specifying an upper bound on the total number of B cells, and allocating the resources r_x (number of B cells) to the ARBs as a function of their stimulation level, *sl* (first normalized in [0,1]), as follows: $r_x = k.(sl^2)$.

Because of the normalization of stimulation, there will always be ARBs left with no resources. These are in turn removed from the network.

Finally, we note that the ARBs are essentially a compression mechanism that takes the B cells to a higher granularity level. However, the population still grows at a prolific rate. We have also observed that the ARB population tends to converge rather prematurely to a state where a *few* good ARBs matching a small number of antigens overtake the entire population because of the resource allocation mechanism.

4. A FUZZY ARTIFICIAL IMMUNE SYSTEM MODEL

4.1 The Balance between Fidelity and Efficiency

In the quest to exploiting all the aspects of the natural immune system within a computational framework, high fidelity to the natural immune system mechanics has been of primordial importance. However, one should not be blind to the practical constraints involved in implementing and later using a computational model for certain applications such as in pattern recognition and data mining. Most organisms' cognitive systems and natural immune systems seem to have almost unlimited ability to learn new concepts. But despite all the latest advances, the computer still has limited capacity compared to the human body. According to the stringent practical demands of today's data mining applications, AINE's information processing and results are still considered extremely expensive both from a computational and storage point of view. Also, the current AIS model leaves plenty of room for needed improvements in terms of the representation of the primary primitives (B-cells and ARBs), as well as for concepts affecting the dynamics of the immune network (stimulation level). We address these points below.

4.2 The Fuzzy Artificial Recognition Ball

The *fuzzy* ARB represents not just a single data item, but instead defines *a fuzzy set* [18] over the domain of discourse consisting of the training data set. The fuzzy set's shape can be any continuous function that decreases with distance from the center of the ARB (prototype / best exemplar). Unlike the original AINE model, each fuzzy ARB is allowed to have its *own* scale / radius of influence (σ_i) which is similar to the *NAT* threshold. However, crisp thresholding is no longer necessary because the fuzzy membership function will gradually exclude antigens that are far away from the prototype. The fuzzy membership function serves also as a *robust* weight function that will decrease the influence of *outliers*. For ARB_i, we define the following weight/membership function:

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{2\sigma_i^2}\right)$$

The *stimulation level* is defined as the density of the antigen population around a certain ARB: $s_i = \frac{\sum_j w_{ij}}{\sigma_i^2} = s_{i\_antigens}$

The above definition has certain desired features from an immune system point of view: the numerator acts like a *stimulation* factor trying to cover as many antigens as possible, while the denominator will limit the covered area of the ARB, thus acting like a *suppression* mechanism. One could also view that the numerator by itself promotes *generalists* (cover as many antigens as possible), while the denominator promotes *specialists* (cover a smaller area). The fraction is expected to achieve a certain *balance* between the two desired extremes. Unlike the original AINE ARB's, each fuzzy ARB scale value (σ_i) is dynamically updated in each iteration to maximize its *stimulation* level (and hence *survival* chances). By setting $\frac{\partial s_i}{\partial \sigma_i^2} = 0$, we obtain $\sigma_i^2 = \frac{\sum_j w_{ij} d_{ij}^2}{\sum_j w_{ij}}$

ARBs located in each other's influence regions should either be merged to limit the population growth or pulled away from each other to explore new areas (for instance by penalizing their stimulation level via a second suppression term). One way to add neighboring ARB *suppression* to explicit antigen stimulation is to redefine stimulation as:

$s_i = s_i$_antigens - $\beta(t)$. s_i_neighboring_antibodies

This will modify the ARB scale update equations to become

$$\sigma_i^2 = \frac{\sum_j w_{ij} d_{ij}^2 - \beta(t) \sum_{ARB_k} w_{ik} d_{ik}^2}{\sum_j w_{ij} - \beta(t) \sum_{ARB_k} w_{ik}}.$$

4.3 The Fuzzy AINE Algorithm

Fuzzy AINE Algorithm:

Initialize fuzzy AINE (ARB pop and σ_i^2) using a cross section of the input data;
Load antigen population = remaining training data;
Repeat Until termination condition {
  Repeat for each antigen
    Present antigen to each fuzzy ARB in network and update w_{ij};
  Repeat for each fuzzy ARB_i {
    Compute ARB_i's stimulation level;
    Update σ_i^2;
  }
  Allocate B cells to fuzzy ARB's based on stimulation level;
  Remove weakest ARBs (0 B cells) from population;
  Clone and mutate remaining fuzzy ARBs;
  Integrate new fuzzy ARBs into fuzzy AINE
}

Note that in the cloning process, the clone inherits the scale value of the parent ARB.

4.4 Avoiding the Premature Convergence of the fuzzy ARB population

The ARBs compete for a finite number of resources (B cells) as specified by an upper bound on the total number of B cells. However, when allocating the resources r_x (number of B cells) to the ARBs as a function of their stimulation level *sl*, we try to limit the influence of the best ARBs to slow their overtaking of the population. This is accomplished by modifying the number of allocated resources, as follows

$$r_x = k.\ sl$$

4.5 Consolidating the final fuzzy ARB population

Algorithm to consolidate the final ARB population:

REPEAT FOR each link (linking ARB_i to ARB_j in Network)
  IF affinity(link) < ε THEN {
    Merge ARB_i and ARB_j into a single ARB_k;
    Aggregate the data contents (antibody) of ARB_i and ARB_j:
      ARB_k.antibody =
        Crossover (ARB_i.antibody, ARB_j.antibody);
  }
}

The crossover of the ARB chromosomes can either be done by combining (randomly exchanging) their, or by any other reasonable aggregation of their data entities, such as arithmetically averaging their attributes using the *'average'* function.

4.6 Controlling the cloning phase

After cloning and mutation, ARBs with *identical* session data information should be merged to limit the population growth.

5. EXPERIMENTS WITH SYNTHETIC DATA

A synthetically generated 2-dimensional set with 600 samples, as shown in Fig. 1 was used to test the algorithm.

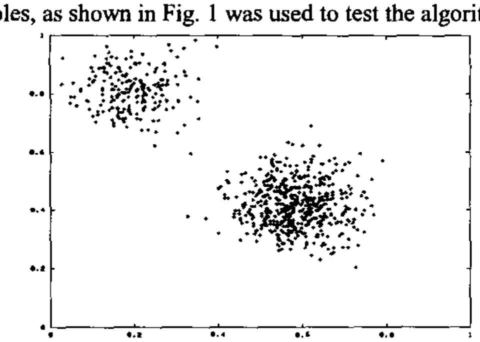

Figure 1: Synthetic data set with two clusters

5.1 Results with original AINE

The parameters of the algorithm were: Maximum Resources =1000, Mutation probability=0.2.

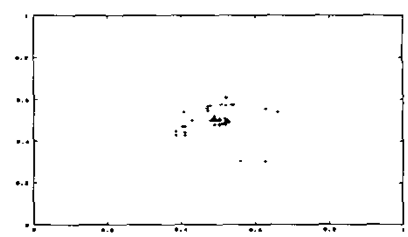

Figure 2: distribution of ARBs in feature space

As seen in Figure 3, the original AINE algorithm was not able to identify the two clusters properly.

5.2 Results with the Fuzzy AINE algorithm

The parameters used were: Maximum Resources =500, Mutation probability=0.2, β=3. The evolution of the ARB population size is shown in Figure 4, and the spatial distribution of the ARB centers at iterations 39 and 49 are shown in Figure 5 and Figure 6, respectively. Note how the population converges and stabilize in a very precise way around the cluster centers.

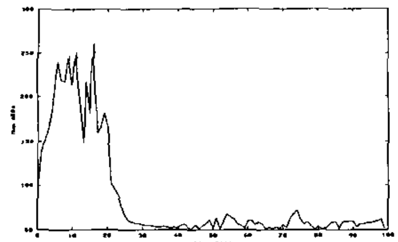

Figure 3: fuzzy ARB population size

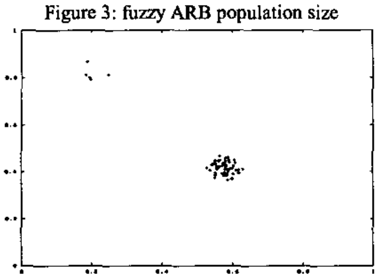

Figure 4: fuzzy ARB distribution (iterat. 39)

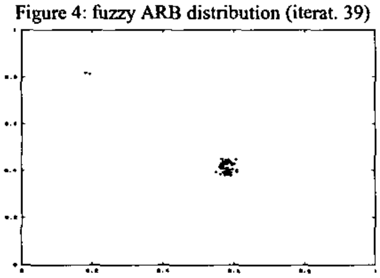

Figure 5: fuzzy ARB distribution (iterat. 49)

Fig. 7 (a)-(f) show the stimulation levels vs. the x-coordinate.

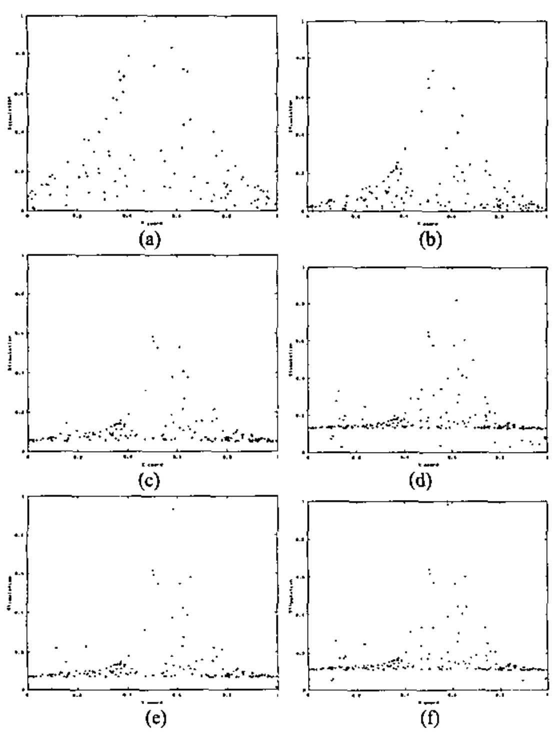

Figure 6:Evolution of fuzzy ARB stimulation levels at iterations 0-5.

Clearly, the stimulation value is higher around the cluster's centers. But the most interesting thing is how initially, the densest (right) cluster dominates, but the ARBs around the other (left) cluster start to increase their stimulation until finally it reaches the same value of the other cluster ARBs. Figure 8 shows the stimulation after several more iterations where the stimulation level of the two clusters become similar and it shows some oscillation.

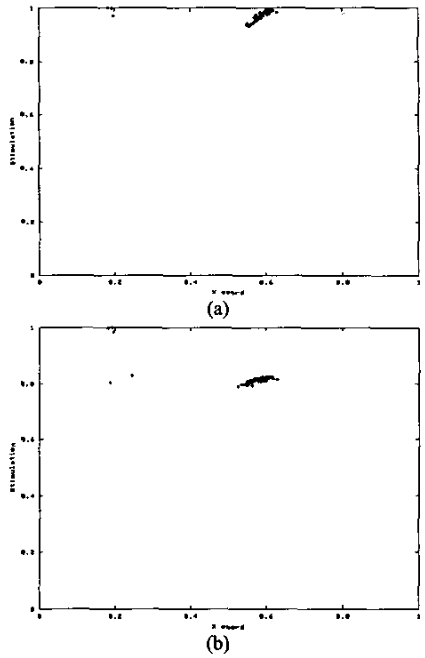

Figure 7: Evolution of the ARB stimulation levels at iterations 49-50.

The explanation of this behavior has to do with the new way to calculate the stimulation levels. Since the suppression term penalizes close ARBs, this mechanism interacts with the cloning mechanism that gives more clones to ARBs with higher stimulation. This prevents very good ARBs from dominating the less good ones, since their stimulation cannot grow beyond a limit. In essence, this is a niching mechanism that is missing in the original AINE, and that prevented a *diverse* ARB population from surviving. The figures also show that the strength of a cluster is reflected by the concentration of ARBs, which is compatible with the natural immune system defense mechanism.

Finally the training data and fuzzy ARBs corresponding to the final iteration (99) are shown in Figure 9, with their sigma values specified by an error bar.

6. LEARNING USER PROFILES USING THE FUZZY AINE MODEL

In our approach, the Web server plays the role of the human body, and the incoming requests play the role of antigens that need to be detected so that specialized antibodies can be learned and added to the antibody repertoire of the simulated immune system. The URL requests have a richer meaning at the granularity of the session level than at the single URL level, because that is one way to distinguish different users or sessions. Hence, the first step in the user profiling consists of preprocessing the Web log data so that it is segmented in different sessions. Each session is then treated as an antigen to be presented to the Web server's artificial immune system.

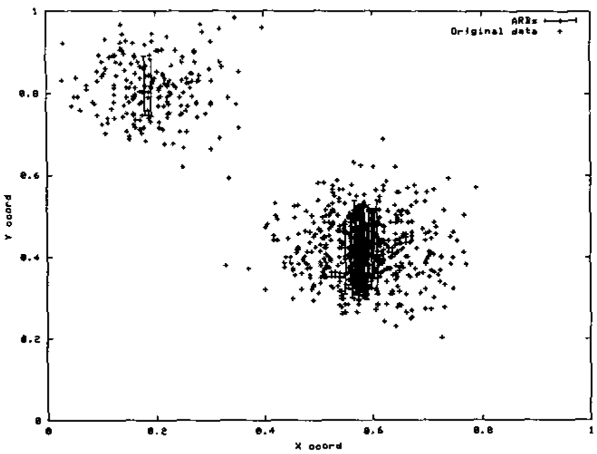

Figure 8: Final ARBs superimposed on original data set.

6.1 Data Preprocessing

The access log of a Web server is a record of all files (URLs) accessed by users on a Web site. Each log entry consists of the following information components: *access time, IP address, URL viewed, ...etc.* An example showing two entries is displayed below

17:11:48 141.225.195.29 GET /graphics/horizon.jpg 200
17:11:48 141.225.195.29 GET /people/faculty/nasraoui/index.html 200

The first step in preprocessing consists of mapping the N_U URLs on a website to distinct indices. A user session consists of accesses originating from the same IP address within a predefined time period. Each URL in the site is assigned a unique number $j \in 1, \ldots, N_U$, where N_U is the total number of valid URLs. Thus, the ith user session is encoded as an N_U-dimensional binary attribute vector $s^{(i)}$ with the property
The features of the session vector are asymmetric binary (a 1

$$s_j^{(i)} = \begin{cases} 1 & \text{if user accessed } j^{th} \text{ URL} \\ 0 & \text{otherwise} \end{cases}$$

value has more weight and meaning than 0). Hence we modify the distance/affinity measure used in the AIS to be based on the *cosine* similarity between two session vectors as

follows: $$S_{kl} = \frac{\sum_{i=1}^{N_U} s_i^{(k)} s_i^{(l)}}{\sqrt{\sum_{i=1}^{N_U} s_i^{(k)} \sum_{i=1}^{N_U} s_i^{(l)}}}$$

Finally, the distance measure is defined as $D_{kl} = 1 - S_{kl}$.

6.2 Learning Usage Trends using the Artificial Immune System

The procedure described in Section 4 is used to present the sessions as foreign antigens to the artificial immune system that will in turn detect strong usage trends and learn them by generating and maintaining a dynamic population of fuzzy Artificial Recognition Balls (ARBs) each representing a family of B cells specialized to detect a particular usage trend. The ARBs form a *network* structure that maintains a *memory* of all past encounters, and that can adapt to new encounters as well. After the final post-merging step of the ARBs, the merged ARBs correspond to typical *profiles* for the users accessing a given website. Note that consolidation of the ARBs is done by averaging their data attributes. The averaged attributes reflect the relevance of the individual URLs to the combined ARBs.

7. WEB USAGE MINING EXPERIMENTAL RESULTS

The fuzzy AINE model was used to detect typical session profiles from the12 day access log data of the Web site of the department of Computer Engineering and Computer Sciences at the University of Missouri, Columbia. After filtering out irrelevant entries, the data was segmented into 1703 sessions accessing 343 distinct URLs. The maximum elapsed time between two consecutive accesses in the same session was set to 45 minutes. The fuzzy AINE algorithm was applied on the training data consisting of the web sessions. The total number of resources (B cells) was limited to 2000, the NAT parameter was set to the average distance in the data set (0.4), and the mutation rate was fixed to 0.02. The parameter Beta was set to 1.0. Figure 10 and 11 show the evolution of the ARB population size and the average population stimulation. Iteration 21 shows a global minimum in the *population size* and global maximum in *average stimulation*, both indicative of a *high-quality* solution. Hence, we chose iteration 21 to extract profiles. Some of these profiles are shown in Table 1. URLs with attributes less than 0.6 (after merging linked ARBs and averaging their attributes) were removed. After the final ARB consolidation step, the individual ARBs' stimulation levels are averaged to yield sl_{avg}. Only 36 profiles had high stimulation levels (≥ 0.7).

8. CONCLUSIONS

Tackling the challenges brought forth by the rapid proliferation and changes both in the structure, content, and usage of the WWW calls for the investigation of innovative together with classical learning techniques. Mining typical user profiles from the vast amount of historical data stored in access logs is an important component of Web personalization. We have introduced several new enhancements to deal with some of the weaknesses of

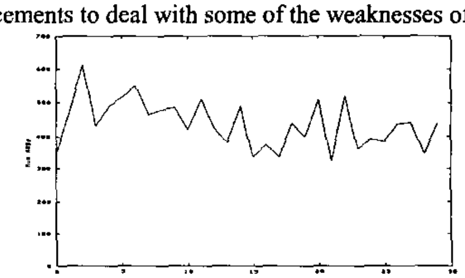
Figure 9: No. of ARBs per iteration

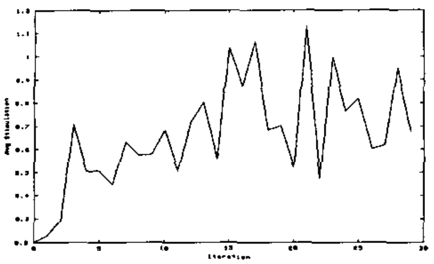

Figure 10: Average population stimulation

TABLE 1. SOME OF THE 36 DETECTED PROFILES

sl_{avg}	Relevant URLs	Profile Description
0.90	/	Main page
0.71	/, /courses.html, /courses_index.html, /courses100.html	Main page and freshman level course list
0.72	/courses400.html, /courses.html, courses_index.html, /courses100.html, /courses_webpg.html, /courses200.html	All level course lists
0.79	/, /people.html, /people_index.html, /faculty.html	Main page, people, and faculty pages
0.90	/, /degrees_undergrad.html, /degrees_index.html, /bsce.html, degrees.html	Main page, and undergraduate degrees
0.95	/~c697168/cecs227/main.html, /faculty/tyrer.html, /~c697168/cecs227 /~c697168/cecs227/head.html /~c697168/cecs227/labs/index.html /~c697168/cecs227/labs/head.html /~c697168/cecs227/labs/left.html /~c697168/cecs227/labs/main.html /~c697168/cecs227/labs/lab1.html	CECS 227 course pages and its course coordinator's homepage
0.71	/~yshang, /~yshang/CECS341.html, /~yshang/W98CECS341	CECS 341 course pages
0.90	/, /degrees.html, /degrees_grad.html, /degrees_grad_index.html, /deg_grad_genor.html	Degrees, graduate degree requirements,
0.95	/~shi/cecs345/java_examples, /~shi/cecs345/Lectures/06.html, /~shi/cecs345/Lectures/07.html,	Specific cecs345's pages

previous artificial immune system models. A fuzzy artificial immune system, mimicking the body's adaptive learning and defense mechanism in the face of invading biological agents, can act as a monitoring and learning system for a Web site in the face of incoming Web requests. Our profiling approach is based on prior traversal patterns of the users, hence, respecting privacy. Also, our immune system inspired approach can be used for different applications ranging from profiling for personalization applications, to dynamic defense mechanism for security maintenance on sensitive e-commerce and general Web sites.

Our immune system approach to detecting different user access patterns borrows its strength from its nature inspired origins. Like the natural immune system, its strongest advantage compared to current approaches is expected to be its ease of adaptation to the changing/dynamic environment that characterizes the World Wide Web. We are currently investigating further enhancements to the AIS model, especially from an efficiency point of view, and plan to investigate its applications in data mining and clustering.

9. ACKNOWLEDGEMENTS

We thank Dr. Jon Timmis of University of Kent at Canterbury, UK, for providing his implementation of the original AINE. D. Dasgupta and F. Gonzalez's work is supported by DARPA (no. F30602-00-2-0514) and NSF (no. NSF-EIA-9818323).

10. REFERENCES

[1] R. Cooley, B. Mobasher, and J. Srivasta, Web Mining: Information and Pattern discovery on the World Wide Web, Proc. IEEE Intl. Conf. Tools with AI, Newport Beach, CA, pp. 558-567, 1997.

[2] D. Dasgupta (editor). *Artificial Immune Systems and Their Applications*. A book published by Springer Verlag Inc., January 1999.

[3] D. Dasgupta. "Artificial Neural Networks and Artificial Immune Systems: Similarities and Differences." In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '97)*, Orlando, October 12-15 1997.

[4] D. Dasgupta and S. Forrest. "Novelty Detection in Time Series Data using Ideas from Immunology. "In *ISCA 5th International Conference on Intelligent Systems*, Reno, Nevada, June 19-21 1996.

[5] D. Dasgupta and N. Attoh-Okine. Immunity-based systems: A survey. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pages 363-374, Orlando, Florida, October 12-15 1997.

[6] L. N. De Castro and F. J. Von Zuben. AiNet: An Artificial Immune Network for Data Analysis. Data Mining: A Heuristic Approach, H. A. Abbass, R. A. Sarker, and C.S. Newton (Eds.), Idea Group Publishing, USA.

[7] S. Forrest, A. S. Perelson, L. Allen, and R. Cherukuri. Self-nonself discrimination in a computer. In Proceedings of the IEEE Symposium on Research in Security and Privacy, IEEE Computer Society Press, Los Alamitos, CA, pp. 202-212, 1994.

[8] S. Forrest, B. Javornik, R. E. Smith and A. S. Perelson. Using genetic algorithms to explore pattern recognition in the immune system. In Evolutionary Computation 1:3, pp. 191-211, 1993.

[9] J. Hunt and D. Cooke. An adaptative, distributed learning system, based on immune system. In *Procedings of the IEEE International Conference on Systems, Man and Cybernetics*, pp. 2494-2499, 1995.

[10] J. Hunt, J. Timmis, D. Cooke, M. Neal and C. King. Jisys: The development of an Artificial Immune System for real world applications, Applications of Artificial Immune Systems, D. Dasgupta Ed., pages 157-186. Pub. Springer-Verlag, 1999.ISBN 3-540-64390-7

[11] J. Hunt and J. Timmis. Evolving and Visualizing a Case Database using an Immune Network. In European Conference on Artificial Intelligence (ECAI 98), 1998.

[12] N. K. Jerne. The immune system. *Scientific American*, 229(1):52-60, 1973.

[13] O. Nasraoui and R. Krishnapuram, and A. Joshi. Mining Web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimator, 8th International World Wide Web Conference, Toronto, pp. 40-41, 1999.

[14] O. Nasraoui, H. Frigui, R. Krishnapuram, and A. Joshi. Mining web access logs using relational competitive fuzzy clustering, 8th International Fuzzy Systems Association Congress, Hsinchu, Taiwan, Aug. 1999.

[15] O. Nasraoui, R. Krishnapuram, H. Frigui, and A. Joshi. Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering, International Journal on Artificial Intelligence Tools, Vol. 9, No. 4, pp. 509-526, 2000.

[16] M. Perkowitz and O. Etzioni. Adaptive web sites: Automatically learning for user access pattern. Proc. 6th int. WWW conference, 1997.

[17] J. Timmis, M. Neal, and J. Hunt. An artificial immune system for data analysis. Biosystems, 55(1/3): 143-150, 2000.

[18] L. Zadeh. Fuzzy sets. Inf. Control 8, 338-353, 1965