Applying Sequence Alignment in Tracking Evolving Clusters of Web-Sessions Data: an Artificial Immune Network Approach

Mozhgan Azimpour-Kivi
School of Engineering and Science
Sharif University of Technology, International Campus
Kish Island, Iran
mojgan_az@kish.sharif.edu

Reza Azmi
Department of Computer Engineering
Alzahra University
Tehran, Iran
azmi@alzahra.ac.ir

***Abstract*—Artificial Immune System (AIS) models have outstanding properties, such as learning, adaptivity and robustness, which make them suitable for learning in dynamic and noisy environments such as the web. In this study, we tend to apply AIS for tracking evolving patterns of web usage data. The definition of the similarity of web sessions has an important impact on the quality of discovered patterns. Many prevalent web usage mining approaches ignore the sequential nature of web navigations for defining similarity between sessions. We propose the use of a new web sessions' similarity measure for investigating the usage data from web access log files. In this similarity measure, in addition to the sequential nature of web navigations, the usage similarity of web sessions is taken into consideration. The ability of the AIS system to track evolving patterns of web usage is validated by applying the proposed method on real world web data.**

Keywords-artificial immune system; web usage mining; web session similarity; sequence alignment

I. INTRODUCTION

The World Wide Web (WWW) is considered as the largest distributed collection of information. By rapid grows of this information resource, it has become a hard task for users to attain their desired information even in a single website. Hence, a need for developing automatic and intelligent client-side or server-side systems that can facilitate this issue has been highlighted. To address the mentioned problem, web usage mining techniques has recently attracted many attentions [1], [2], [3], [4], [5]. Web usage mining techniques tries to extract interesting patterns from the data that are collected from the interaction of users with the web. The aim of any web usage mining process is to learn models of users' behavior and apply these models for any application that tries to ease use of the web. These applications include creating adaptive websites, web personalization, web recommender systems, etc. Web has some inherent properties that can affect designing of any system that deals with the web. Since the content of the web is very dynamic, the users' navigational behaviors also changes gradually. Moreover, periodic deletion and insertion of contents on the web makes it a noisy source of information. Finally, the web is huge, thus a massive amount of web data is usually needed to discover all the users' behavior patterns. Considering these properties, any web usage mining system that is used to discover the web should be adaptive, robust to noise and also scalable.

The Artificial Immune System (AIS) is a powerful paradigm for learning which is originally inspired from the natural immune system. A special type of white blood cells (called B-cells), are responsible for detecting antigens (such as viruses) and defending against them in vertebrate immune system. When an antigen is detected by the B-cells, an immune response is promoted resulting in antigen elimination. One type of response is the secretion of antibodies by B-cells (cloning). Antibodies are Y-shaped molecules on the surface of B-cells that can bind to antigens and recognize them. Each antibody can recognize a set of antigens which can complementarity match the antibody. The strength of the antigen-antibody interaction is measured by the affinity of their match. For the sake of simplicity, we do not differentiate between antibodies and B-cells.

Artificial immune system has many outstanding features which make it attractive for applications that cope with evolving data [6]. These properties include: recognition, diversity, self regulation, robustness and memory. Note that, many conventional web usage mining techniques, such as data mining techniques, should handle very large datasets. Even if the techniques are scalable, they are not usually robust to the noise which is the nature of the web data. Moreover, they are unable to mine evolving web data.

Many artificial immune models have been discussed in literature such as Negative Selection, Danger Theory and Artificial Immune Networks (AINs). Of particular relevance to our work, is the AIN model which was initially proposed by Jern [7]. The main idea of this method is to regulate the immune system by antibodies which are interacting to each other in a network graph structure. The resource limited Artificial Immune System (RLAIS) model, proposed by Timmis and Neal [8], promote more general data analysis by introducing the concept of ARBs (Artificial Recognition Balls). Each ARB is consists of several identical B-cells, which could match by Euclidean distance to an antigen or to another ARB in the network. Each member of the antigen in training set is matched against each ARB based on the Euclidean distance. This affects ARB's stimulation level which is inversely related to its average distance from the antigens. ARBs compete for allocating resources from the resource pool based on their stimulation level. They also can interact to each other and clone themselves based on their

978-0-7695-4482-3/11 $26.00 © 2011 IEEE
DOI 10.1109/CICSyN.2011.22

IEEE computer society

stimulation level. Nasraoui et al. [2] proposed a fuzzy model of ARBs based on RLAIS for web usage mining task. The Fuzzy ARB represents not just a single data item, but instead defines a fuzzy set. Each fuzzy ARB has its own radius of influence which decrease as its distance to another data point is increased. In [3], a new immunity inspired approach is presented called TECNO-Streams (Tracking Evolving Clusters in NOisy Streams). In [3] and [9], this system is used for tracking evolving clusters in web sessions data.

One of the challenging issues in analyzing web session data is defining a measure of similarity between two web sessions. A more precise similarity measure can definitely present the nature of data properly. Euclidean distance, Cosine similarity, and Jaccard coefficient are the most popular measures that are used for this goal. Indeed, a web session contains a sequence of URLs accessed by a user. Therefore, a good similarity measure should be defined so that it does not ignore the sequential nature of web navigations in sessions. On the other hand, not all of the URLs visited in a session are equally important to the users. Current immunity inspired web usage profiling systems represent a session using a vector defined over the space of web pages within a particular website [3], [4], [5]. This data representation has a high memory usage. In addition, Euclidean distance or Cosine similarity which is usually applied for estimating the similarity of two session vectors ignore the sequential nature of web navigations.

Previously, we proposed a web sessions similarity measure using sequence alignment method [10]. In this similarity measure, not only the sequential nature of web navigations is considered, but also the usage similarity of two web sessions is taken into account. The usage similarity of web sessions is defined based on the time a user spends on a webpage and the frequency of visitation of each page within the session. In this case, we represent a session by a sequence of URLs that are accessed by a specific user. The goal of this paper is to apply the mentioned similarity measure for generating an AIN of B-cells (web sessions) using TECNO-Streams. Furthermore, due to the change in data representation, some modifications are necessary in TECNO-Stream procedures regarding to the network compression strategy [3]. The ability of the new AIS is evaluated by applying the proposed method on some real world web sessions data. The results confirm the ability of the proposed system in tracking emerging clusters of web sessions and identifying the similar groups of web navigations continuously. Owed to the fact that the sequential nature of web navigations is not discarded, we believe that the discovered patterns can represent the nature of web sessions data more properly.

The remainder of this paper is organized as follows. In Section 2, the TECNO-Streams algorithm is reviewed. Section 3 discusses the goals of this study and introduces some modifications in TECNO-Streams algorithm regarding the data representation and the network compression method. In Section 4, the similarity measure of web sessions using sequence alignment is described briefly. Section 5 describes how we can apply the proposed system for tracking evolving clusters of click stream data and the evaluating results are presented. Finally, Section 6 concludes our study.

II. TECNO-STREAMS (TRACKING EVOLVING CLUSTER IN NOISY STREAMS)

In this section we briefly review the features of TECNO-Streams that are relevant to our work, leaving most of the details in [3]. TECNO-Streams is composed of a set of Dynamic Weighted B-cells (D-W-B-cells) that tend to summarize the learned model. In addition, there are some stimulating and suppressing interactions between D-W-B-cells. The learning process starts by presenting a set of input data (antigens) to the network of B-cells one at a time. The system tries to learn an optimal network of linked D-W-B-cells using cloning operation, as described in [3]. Each D-W-B-cell represents a learned pattern that can match to an antigen or another D-W-B-cell. In addition, each D-W-B-cell represents a softly defined influence zone that is described in a term of weight function which decreases with distance from the antigen and the time since the antigen has presented to the network. The strength of the link between two D-W-B-cell is directly related to their similarity. This provides a co-stimulation between similar B-cells that support them even in the absence of the antigen that caused their creation. Note that the co-stimulation may cause explosive growth of the B-cells population in a local neighborhood. To control the redundancy issue in the network a co-suppression phenomenon is also regarded in the B-cells interactions.

The activation of i^{th} D-W-B-cell caused by j^{th} antigen in the network after J antigen are presented to the network is defined by (1). In this equation, d_{ij}^2 is the distance from antigen j to D-W-B-cell i. σ_{ij}^2 is the scale factor that defines the size of the influence zone around a cluster prototype. τ is a constant that determines the rate of forgetting in immune network.

$$w_{ij} = e^{-\left(\frac{d_{ij}^2}{2\sigma_{ij}^2} + \frac{(J-j)}{\tau}\right)} \tag{1}$$

The stimulation level of a D-W-B-cell after presenting J antigen to the network and the optimal scale of i^{th} D-W-B-cell can be calculated based on (2) and (3) respectively. Consider that, these two equations can be updated by using incremental approximations described in [3].

$$\delta_{ij} = \frac{\sum_{j=1}^{J} w_{ij}}{\sigma_{ij}^2} + \alpha(t)\frac{\sum_{l=1}^{N_B} w_{il}}{\sigma_{ij}^2} - \beta(t)\frac{\sum_{l=1}^{N_B} w_{il}}{\sigma_{ij}^2} \tag{2}$$

$$\sigma_{ij}^2 = \frac{\sum_{j=1}^{J} w_{ij} d_{ij}^2}{2\sum_{j=1}^{J} w_{ij}} \tag{3}$$

The first term on the right side of (2) describes the pure stimulation of D-W-B-cell caused by antigen *j*. Also, the second and third terms represent co-stimulation and co-suppression interactions from other B-cells in the network respectively. The parameter N_B is the maximal number of D-W-B-cells in the network. The parameters $\alpha(t)$ and $\beta(t)$ are stimulation and suppression coefficient of D-W-B-cell and are updated based on the age (t) of the B-cell. If an antigen could stimulate the B-cell sufficiently, ($w_{ij} \geq w_{min}$), then the age of this B-cell is refreshed to zero. Otherwise, it increases by one. The coefficients increase as the age of the B-cell decreases, hence recently activated B-cells have more impact on the network.

Nasraoui et al. [3], propose to use a K-means clustering algorithm periodically to group almost similar B-cells. Then, the internal (B-cell to B-cell) interactions of the B-cells can be considered only in the most activated group of B-cells in the network. The most activated group is the group which its centroid has the largest weighted distance (w) to presented antigen. Applying this method reduces the number of total integrations that should be calculated in the network greatly.

III. PROPOSED METHOD

In this paper, we tend to use the basic concepts of TECNO-Streams to track evolving clusters in web sessions data. As described earlier, the definition of similarity measure between sessions can have a great impact on the quality of the discovered patterns. As mentioned, we tend to use a similarity measure that considers not only the sequential nature of web navigations, but also the interestingness of visited web pages to the users.

Consider that, the data representation of D-W-B-cells and the antigens in the system should change from binary representation. In this case, we represent web sessions as a sequence of pages which a user has visited them. Also, we need to have the corresponding time and the frequency of visitations of each web page in a session in order to estimate the interestingness of each web page to the user. Reduction in memory usage by applying the introduced data representation of web sessions is an advantage of this method. By applying the discussed change, the network compressions strategy (K-means clustering algorithm), discussed in previous section, is not applicable in the AIS. As the new data points are categorical data, we cannot get average of the attributes of data points belonging to a particular cluster to calculate the centroids of K-means clustering. We propose to use an alternative solution using the concept of K-Nearest Neighbor (KNN). In this case, every B-cell knows its KNN in the network. When an antigen is encountered to the network, the activated B-cells and their KNNs are considered as a sub-network of almost similar B-cells. In other words, the internal interactions of B-cells in this group are only calculated by considering the members of this sub-network. Thus, the total number of internal interactions of B-cells reduces greatly. The algorithm for the proposed method is presented in Fig. 1.

1- Fix the Maximal population size N_B;
2- Initialize D-W-B-cell population and $\sigma_i^2 = \sigma_{init}$ using a number of random antigens;
3- Repeat for each antigen
    1. Present antigen to each D-W-B-cell;
    2. If antigen activated the D-W-B-cell ($w_{ij} \geq w_{min}$)
        i. Refresh age ($t = 0$);
        ii. Add the current D-W-B-cell and its KNN to working sub-network;
    3. Else
        i. Increment the age of D-W-B-cell by one;
    4. If for all D-W-B-cells $w_{ij} \leq w_{min}$
        i. Create a new D-W-B-cell = antigen_j;
    5. Else
        i. Repeat for each D-W-B-cells in working sub-network;
            1. Compute D-W-B-cell stimulation using (2);
            2. Update D-W-B-cell σ_i^2 using (3);
    6. Clone D-W-B-cells based on their stimulation level;
    7. If population size > N_B;
        i. Remove extra least stimulated D-W-B-cells

Figure 1. The modified algorithm of TECNO-Streams.

As presented in Fig. 1, when an antigen is unable to activate any B-cell, this antigen may represent a noise or a new emerging pattern. In this condition, a new B-cell is created which is a copy of the presented antigen. If this antigen is a noisy data and does not present a new emerging pattern, it would not get enough chance to get stimulated by incoming antigens and is probably eliminated. After each antigen is presented to the network, the B-cells go under cloning operation based on their stimulation level as described in [3]. Note that due to our data representation, we can randomly replace a URL in a clone of a B-cell by another valid URL of the website under study. When the population of the network exceeds a defined threshold, the least stimulated B-cells are removed from the network.

The distance measure presented in this study is used in all the steps for calculating the internal and external (B-cell to antigen) interactions of B-cells. The detailed information of this similarity measure is described by the authors in [10]. In the following section, we only have a brief review on it.

IV. THE SIMILARITY MEASURE OF WEB SESSIONS USING SEQUENCE ALIGNMENT

In the proposed method for calculating the similarity of web sessions, not only the sequential order of visited web pages is taken into consideration, but the usage similarity of two web sessions is considered. First, the similarity of two web pages in two sessions is defined based on the conjunction of the similarity of the web pages' URLs and the similarity of their interestingness to users. Then, the sequence alignment method [11] is applied in order to find the best match of two sessions and to estimate the similarity of two session sequences.

For finding the similarity of two web page URLs we utilize the method proposed by Wang and Zaiane [12]. This method does not consider the content of web pages but simply the paths leading to a webpage in the hierarchical structure of the URLs of the website. The detail information of this method is discussed in [12]. Briefly, each level of URLs is represented by a token. Then for estimating the similarity of two URLs, based on the length of the longer token string, a decreasing weight is assigned to each level of tokens from the last to the first. Finally, the similarity between two strings of tokens (*Token_Sim*) is defined as the sum of the weights of matching tokens divided by the sum of the total weights. As mentioned earlier, in addition to the similarity of two web page URLs, the similarity between the interests of their users in visiting those pages (*Interest_Sim*) is important. The interestingness of web pages in sessions is defined based on the harmonic mean of the normalized spent time on pages and frequency of visitation from those pages as described in [10]. Considering the *Token_Sim* and *Interest Sim*, the similarity of two web pages P_i and P_j can be defined as shown in (4). In this equation, α is a scale factor which defines the contribution of *Token_sim* and *Interest_sim* in final similarity of web pages. This parameter should be assigned with a value between 0 and 1.

$$\begin{aligned}\text{Similarity}\ (P_i, P_j) = \alpha\ \text{Token}\_\text{Sim}(P_i, P_j) \\ + (1-\alpha)\ \text{Interest}\_\text{Sim}(P_i, P_j)\end{aligned} \quad (4)$$

For estimating the similarity of two web session sequences, we apply the sequence alignment method in order to find the best match between two sequences. For applying the sequence alignment method, we need to define a scoring function which helps find the optimal matching between two session sequences. Consider that, the similarity of two web pages discussed in previous section plays the role of a page matching goodness function. The scoring function deployed in the method of this paper is as follows. For each identical matching, i.e. a pair of pages with similarity 1, the similarity score is 20; for each mismatching, i.e. a pair of pages with similarity 0, or matching a page with a gap, the similarity score is -10; for a pair of pages with similarity between 0 and 1, the score for their matching is between -10 and 20. Hence, the scoring function for the similarity between two pages P_i and P_j can be calculated using (5).

$$\text{Score}\ (P_i, P_j) = -10 + 30 \times \text{Similarity}(P_i, P_j) \quad (5)$$

As mentioned earlier, the estimation of the similarity between two web sessions is calculated using the sequence alignment method in order to find the best match between two sequences. This method is facilitated using dynamic programming. The final similarity between the two sequences is obtained based on their optimal matching and the length of the sequences.

V. EXPERIMENTAL EVALUATION

The empirical evaluation reported in this paper is performed on a group of web sessions collected from the Music Machine website by Perkowitz and Etzioni [13]. The original data used in our experiment, contains 62668 requests from log files of the web server for five random days in 1998. After pre-processing task on the original data using methods described in [14], 1233 web sessions were extracted. We applied the agglomerative clustering using average linkage method [1] on the matrix of pairwise distances of web sessions and found 10 interesting profiles of web users visiting this web site. Some of these usage trends are reported in Table 1.

The maximal population size of the network is set to 100, the control parameter for the number of nearest neighbors (K) is set to 5. The activation threshold (w_{min}) is 0.6 and τ = 100. We illustrate the continuous learning ability of the proposed system, using two different scenarios. In the first scenario, we assign the web sessions to the closest profile investigated using agglomerative clustering. Then we present these sessions, ordered by their assigned profiles, one profile at a time to modified TECNO-Streams. In the second scenario, we assume we have the knowledge of session partitions as scenario 1, but we present the sessions by their chronological ordered as recorded by the web server. We track the B-cells that are successful in learning each of 10 basic profiles by checking the existence of at least one B-cell which its weighted distance to the profiles is greater than θ. When there is at least one B-cell in the network that is sufficiently close to one of the ground profiles, a hit is recorded in the system. The evolving number of hits per profile for both of above scenarios is illustrated in Fig. 2 and Fig. 3 respectively. In these figures, a hit for i^{th} profile for session number j is represented by marking the position (i, j). The distribution of input sessions over the ground profiles when sessions are presented by order according to their assigned profiles and when they are in their natural order are also depicted in Fig. 4 and Fig. 5.

TABLE I. SUMMERY OF SOME USAGE TRENDS DISCOVERED USING AGGLOMERATIVE CLUSTERING OF WEB SESSIONS USING AVERAGE LINKAGE METHOD

Profile number	Profile Summery
1	{'/categories'}, {'/categories/effects'}, {'/categories/effects/Alesis'}, {'/categories/patchable'}, {'/categories/patchable/2600'}
6	{'-'}, {'/guide'}, {'/guide/icons.html'}, {'/guide/finding.html'}, {'/guide/structure.html'}
7	{'-'}, {'/links'}, {'/links/misc.html'}, {'/links/machines.html'}
8	{'-'}, {'/manufacturers'}, {'/manufacturers/Roland'}, {'/manufacturers/Roland/Juno/patches'}
9	{'/gearlists'}, {'/gearlists/Dubtribe'}, {'/gearlists/Xymox'}, {'/gearlists/devo-gear'}, {'/gearlists/orb'}

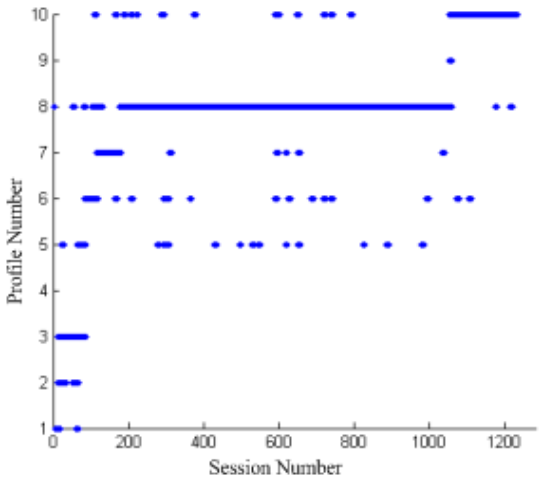

Figure 2. Hits per usage profiles when sessions are presented in order of trends 1 to 10 and θ = 0.7.

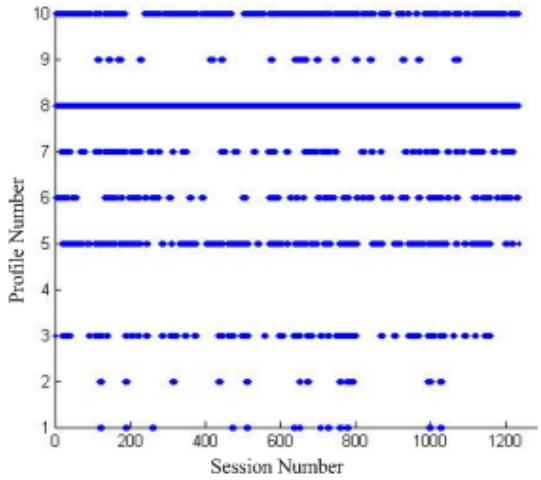

Figure 3. Hits per usage profiles when sessions are presented in their natural order of trends and θ = 0.65.

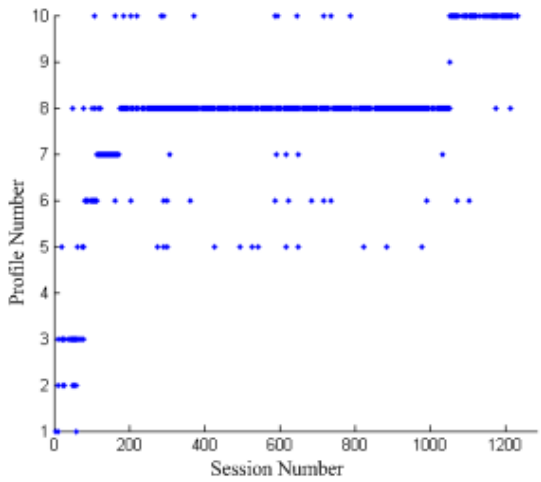

Figure 4. Distribution of input sessions over usage profiles when they are presented in order to their usage trends and θ = 0.7.

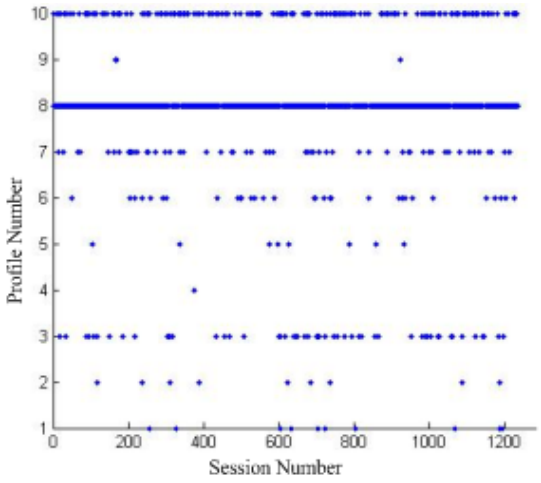

Figure 5. Distribution of input sessions over usage profiles when they are presented in their natural order and θ = 0.65.

Fig. 2 presents an almost staircase pattern which indicates the gradual learning of usage profiles as they are encountered to the immune network in order of their usage profiles. The plot shows some unexpected behavior. As an example, we expect the trends 6 and 7 to be exist for sessions with number around 100 to 200. Yet, we can see these patters in sessions with other numbers. This issue may be caused by a certain tolerance allowed for the inexact matching contributed by parameter θ in detecting hits per profile.

Fig. 3 represents the evolving number of hits per usage profile when sessions are presented in their natural order to the system. According to this figure, the trends 1, 2 and 4 are weak. The distribution of input sessions over usage profiles represented in Fig. 5, indicates that in real world the arrival sequence of different usage trend may be quite surprising and varies greatly with time.

Comparing Fig. 2 to Fig. 4 and Fig. 3 to Fig. 5, we can notice the persistence of a particular pattern in the immune network for a period of time (memory), since the lines for the B-cells are lightly bolder than that of web sessions. As mentioned, the memory span of the network depends to the value of τ, as a higher value provides slower forgetting. Note that the patter shown in Fig. 2 should be very close to that of Fig. 4. Also, we expect the pattern of Fig. 3 to be very similar to that of Fig. 5. As illustrated in these figures, these expectations are closely met. This fact indicates the ability of the proposed system in making a dynamic synopsis of the usage data.

An important advantage of using the proposed similarity measure is that the discovered profiles from the web data can be defined so that they represent a sequence of web navigations which are interesting to the users. For instance, these profiles can be represented by B-cells that are activated more than a defined threshold. Consider that, B-cells would keep the sequential order of web pages in sessions.

VI. CONCLUSIONS

In this paper, we proposed some modifications to the TECNO-Streams AIS for web usage profiling regarding to the data representation of this system and also the network compression method. The method proposed in this study considers the sequential nature of web navigation in web sessions for estimating the similarity of them. In addition to the similarity of web pages' URLs, the usage similarity of web pages visited in sessions is considered for defining the similarity of web pages. As we discussed, we represent a session in the immune system as a sequence of URLs that are requested from the web server. By using the new data representation, we need a new method for compressing the network and reducing the number of immune cells interactions calculated in the AIS. As we described, we utilized a KNN approach to find close population of B-cells to the antigens presented to the network.

The evaluation was performed using some real world web sessions. The sessions obtained from the pre-processed log files of web server are presented to the system as antigens. The network of the B-cells represents a summarized version of the antigens encountered to the network. Also, they are able to adapt to emerging usage patterns proposed by new antigens at any time. The results show the ability of the introduced AIS to track different usage patterns as they are presented to the network. Also, the mined profiles keep the information regarding to the sequential nature of web sessions. The limitation of the proposed method is the high time complexity of sequence alignment methods which its reduction is an open research area.

REFERENCES

[1] T. Hussain, S. Asghar, and S. Fong, "A hierarchical cluster based preprocessing methodology for Web Usage Mining," in *6th Intl. Conf. on Advanced Information Management and Service (IMS)*, Seoul, South Korea, 2010, pp. 472–477.

[2] O. Nasraoui, F. Gonzalez, and D. Dasgupta, "The fuzzy artificial immune system: motivations, basic concepts, and application to clustering and Web profiling," in *Proc. of the 2002 IEEE Intl. Conf. on Fuzzy Systems, 2002. FUZZ-IEEE'02*, Honolulu, HI, USA, pp. 711-716.

[3] O. Nasraoui, C. C. Uribe, C. R. Coronel, and F. Gonzalez, "TECNO-STREAMS: Tracking Evolving Clusters in Noisy Data Streams with a Scalable Immune System Learning Model," in *3rd IEEE Intl. Conf. on Data Mining (ICDM'03)*, Los Alamitos, CA, USA, 2003, vol. 0, p. 235.

[4] O. Nasraoui, C. Cardona, C. Rojas, and F. Gonzalez, "Mining evolving user profiles in noisy web clickstream data with a scalable immune system clustering algorithm," in *Proc. of WebKDD*, Washington, DC, USA, 2003, pp. 71–81.

[5] B. H. Helmi and A. T. Rahmani, "An AIS algorithm for Web usage mining with directed mutation," in *IEEE Congress on Evolutionary Computation, 2008. CEC 2008 (IEEE World Congress on Computational Intelligence)*, Hong Kong, China, 2008, pp. 3122-3127.

[6] T. Ren, Z. Gao, W. Kong, Y. Jing, M. Yang, and G. M. Dimirovski, "Performance and robustness analysis of a fuzzy-immune flow controller in ATM networks with time-varying multiple time-delays," *Journal of Control Theory and Applications*, vol. 6, no. 3, pp. 253-258, Sep. 2008.

[7] N. K. Jeme, "Towards a network theory of the immune system," *Ann Immunol (Paris)*, vol. 125, no. 1-2, pp. 373–389, 1974.

[8] J. Timmis and M. Neal, "A resource limited artificial immune system for data analysis," *Knowledge-Based Systems*, vol. 14, no. 3-4, pp. 121–130, 2001.

[9] O. Nasraoui, C. Rojas, and C. Cardona, "A framework for mining evolving trends in Web data streams using dynamic learning and retrospective validation," *Computer Networks*, vol. 50, no. 10, pp. 1488-1512, Jul. 2006.

[10] M. Azimpour-Kivi and R. Azmi, "A Webpage Similarity Measure for Web Sessions Clustering Using Sequence Alignment," in *Intl. Symp. on Artificial Intelligence and Signal Processing (AISP2011)*, Tehran, Iran, 2011, in press.

[11] K. Charter, J. Schaeffer, and D. Szafron, "Sequence alignment using FastLSA," in *Intl. Conf. on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS'2000)*, 2000, pp. 239–245.

[12] W. Wang and O. R. Zaiane, "Clustering Web sessions by sequence alignment," in *DEXA '02: Proc. of the 13th Intl. Workshop on Database and Expert Systems Applications*, Washington, DC, USA, 2002, pp. 394-398.

[13] M. Perkowitz and O. Etzioni, "Adaptive sites: Automatically learning from user access patterns," in *Proc. of 6th Intl. World Wide Web Conf., Santa Clara, California*, 1997.

[14] V. Sathiya Moorthi and V. Murali Bhaskaran, "Data preparation Techniques for Web Usage Mining in World Wide Web–an approach," *International Journal of Recent Trends in Engineering*, vol. 2, no. 4, 2009.