An Artificial Immune System Approach to News Article Recommendation

Branko Mihaljević, Igor Čavrak and Mario Žagar
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb. Croatia
branko.mihaljevic@fer.hr, igor.cavrak@fer.hr, mario.zagar@fer.hr

Abstract. *Artificial immune systems are solution finding techniques often used for classification and recommendation problems. Danger theory is one of new context dependant response theories of how an artificial immune system responds to pathogens.*

News articles recommendation systems solve problems of presenting articles with interesting topics to user honoring evolving user preferences and past choices. This paper describes how artificial immune system with Danger theory can be utilized for news articles recommendation on Web portals or similar media presenter systems and presents algorithm and method for handling user preferences and article features in recommender system.

Keywords. Artificial immune system, Danger theory, recommender system, news article recommendation.

1. Introduction

Artificial immune systems (AIS) emerged in the last ten years as a new computational paradigm in artificial intelligence. They belong to a group of biologically inspired systems like artificial neural networks, DNA computations, evolutionary algorithms, but they have also been compared to other soft computing paradigms like fuzzy systems and probabilistic reasoning. Independently or in cooperation with other approaches they have already proved their capabilities, but new areas of applications as well as improvements arise every day.

Immune system in general is a complex of cells, molecules, and organs proven to be capable of performing tasks like pattern recognition, learning, self organization, memory acquisition, generation of diversity, noise tolerance, anomaly detection, generalization, distributed detection, and optimization. Artificial immune systems are defined as *adaptive systems, inspired by theoretical immunology and observed immune functions, principles and models, which are applied to problem solving* [1]. Newly developed computational techniques based on immunological principles and immune engineering concepts are used for solving computational problems using metaphors from innate immune systems and self/non-self discrimination. They have properties like adaptivity, diversity, robustness, distributivity, predator-prey pattern, etc.

Applications of AIS approach include many different areas [2] like computational security, anomaly detection, optimization, fault diagnosis, robotics, machine learning, pattern recognition, etc. Many of these areas are covered with hybrids of AIS and other approaches, but some techniques use AIS as an improvement for other algorithm imperfections.

A Danger theory (DT) as an addition to AIS explains certain phenomena in immune processes with the main idea that immune system responds to danger signals, not just the presence of perceived non-self. This theory has been used in anomaly detection (fault detection) and classification (data mining, Web mining) [3].

News articles recommendation problem is common in personalized dynamic media information sources e.g. Web portals. Areas of application could include online forums, blog searches, online help searches, and site content recommendation. The main idea behind this paper relies on the fact that AIS+DT technique can be applied for construction of relatively simple recommender system with properties of constant learning from user interactions and adapting to changeable user preferences.

2. Natural immune system

Multilayered natural immune system is responsible for protection of the human body from foreign invaders by differentiating self from non-self and neutralizing dangerous pathogens and toxins. The immune system consists of many physical, physiological, biochemical, and cellular

barriers, which make the first line of defense against several types of microorganisms. This unchanging mechanism called *innate immune system* detects and destroys certain invading organisms. In addition, *adaptive immune system* responds to previously unseen foreign cells. Accordingly, antibody production in response to specific infections is called *adaptive or specific immune response*. Cells and molecules of immune system recognize almost unlimited variety of infectious foreign cells and substances.

2.1. Basic elements

Lymphoid tissues and organs are responsible for production and maturation of *lymphocytes* [4]. Immune cells are structurally divided into lymphocytes and phagocytes, granulocytes, and their relatives. *B and T lymphocytes* mediate the adaptive immune response and are responsible for the recognition and elimination of pathogenic agents. Main function of *B lymphocytes* is production and secretion of antibodies, specific proteins that recognize and bind to other particular proteins called antigens. *Antigens* are found on the surface of invading cells and the binding of an antibody to an antigen activates elimination. Three types of *T lymphocytes* regulate other cell actions including defensive response. T helper (T_H) cells have a role in activation of B cells, other T cells, natural killer cells, and macrophages. Cytotoxic T killer (T_K) cells eliminate microbial invaders, cancerous cells, and viruses. Suppressor T cells maintain and control immune response preventing autoimmune and allergic reactions.

2.2. Immune processes

Adaptive immune response incorporates features like sufficient diversity to deal with all kinds of antigens, long lasting immunological memory and discrimination of self from non-self.

B and T lymphocytes carry surface receptor molecules capable of recognizing antigen with distinct characteristics. B cells receptors (BCR), called *antibodies* or immunoglobulin, and epitopes of an antigen to which they bind must be complementary. Binding strength between antigen and antibody is called *affinity*. *Cross-reactivity* term describes recognition of many similar antigen epitopes by single sufficiently similar lymphocyte receptor. Before exposing B cells to antigen, combinatorial recombination creates population of cells that vary widely in their specificity.

T cells receptors (TCR) recognize antigens by major histocompatibility complex (MHC) surface molecule bonded to particle of antigen named peptide, left after digestion of an antigen by antigen presenting cell (APC). T cell whose TCR recognizes MHC/peptide secretes lymphokines as mobilization signals for other cells. Activated B cells divide and differentiate into *plasma cells* secreting antibody proteins which neutralize antigens or precipitate their destruction.

2.2.1. Clonal selection

Clonal selection theory describes a basic method of an immune response to antigenic stimulus [5]. B lymphocyte cell with good binding to antigen responds to the antigen stimulus by producing only one kind of antibodies. When B cell receptors bind to antigen, a second signal from T_H cell stimulates proliferation (cloning) of B cells and maturing these new clones into plasma cells. Plasma cells do not proliferate, but are active antibody secretors in contrast to proliferating B cells. Clonal selection or expansion theory explains that only B cells with closer match to antigen will be largely reproduced, much more than clones with lower affinity. Similar process is done with T cells, but the main difference is that B cells undergo somatic *mutation* during reproduction, which increases repertoire diversity. B cells differentiate into long-lived *memory cells*, which do not produce antibodies, but serve in plasma cells production when exposed to second antigenic stimulus. Self-reactive cells are eliminated using *negative selection*, before inducing autoimmune response.

Learning and *memory* principles of immune reactions are acquired through vaccination and prior diseases, causing shorter lag time, higher rate of antibody production, and longer persistence of antibody synthesis in secondary or cross-reactive response. *Positive selection* is a process of eliminating useless lymphocytes by rescuing most efficient cells from cell death, thus controlling survival and differentiation of repertoires.

2.2.2. Idiotypic network theory

Idiotypic network theory introduced by Jerne [6] suggests that interactions in the immune

system do not occur just between antibodies and antigens, but that all components of the network may interact with each other. *Epitope* is a unique shape located on antigen's surface, and *paratope* is a part of antibody responsible for recognizing complementary epitopes. *Idiotype* is a set of shapes, called *idiotopes*, which serve in binding of two antibodies. An antibody may be matched and bonded by other antibodies through paratope-idiotope binding, similar to paratope-epitope binding. Antibodies' bindings spread through the population with positive or negative effects on each particular antibody production. This hypothesis is disputed by some immunologists, but it explains features like repertoire selection, self/non-self discrimination, tolerance, and memory.

3. Artificial immune system model

Artificial immune system is build upon a framework consisting of three basic elements: a component representation, a set of functions quantifying interactions between components, and procedures governing system dynamics. Components could be represented as B-cell, T-cell, antibody, or antigen objects. Affinity functions define similarity measure used for recognition. Procedures are general purpose algorithms used for immune processes analogies. Elements must be chosen prior to modeling regarding to application domain and problem.

3.1. Shape spaces and affinity measures

Generalized shape of a molecule with a set of *L* parameters is presented as a feature vector in a L-dimensional *shape space S* [1]. Shape spaces can be divided to: integer, real-valued, Hamming (attribute string built out of a finite alphabet or enumerations), symbolic, and other shape spaces. *Degree of match* or *affinity* is nonnegative real number which corresponds to distance measure functions $S^L \times S^L \to \Re^+$ between molecules.

Euclidean distance of antibody and antigen is given by Equation (1).

$$D = \sqrt{\sum_{i=1}^{L} (Ab_i - Ag_i)^2} \qquad (1)$$

Binary shape space is a type of Hamming shape space with Hamming distance measure given by Equation (2).

$$D = \sum_{i=1}^{L} \delta_i, \quad \delta_i = \begin{cases} 1 \; if \;\; Ab_i \neq Ag_i \\ 0 \quad otherwise \end{cases} \qquad (2)$$

Two molecules can interact in multiple possible alignments, and there are many different similarity measures e.g. r-contiguous and multiple contiguous bit rule, or affinity measure of Rogers and Tanamoto. *Cross-reactivity threshold* ε describes a recognition or activation threshold that has to be crossed for positive recognition of two molecules, i.e. $D \geq \varepsilon$.

3.2. Algorithms and processes

Several general purpose immune algorithms model specific processes of the immune system. Two major cell generation algorithms are *bone marrow model*, used to generate repertoire of L-dimensional objects with attribute string using pseudo random number generator, and *thymus model* used to generate and differentiate repertoire of cells capable of performing pattern recognition, i.e. self/non-self discrimination. Thymus model uses *positive selection algorithm*, where only cells with affinity greater than the cross-reactive threshold are positively selected and introduced to the system (survival). It also uses *negative selection algorithm* to eliminate cells capable of binding with self. If affinity of an immature cell to at least one self-peptide is greater than the cross-reactive threshold, the cell is eliminated from the system (death).

Two distinct theories are presented in clonal selection algorithm and immune network algorithm. *Clonal selection algorithm* CLONALG presented by de Castro and Von Zuben [7] includes *clonal selection* (selection of elements with the highest affinity), *clonal expansion* (cloning in proportion to the affinity), *affinity maturation* (mutation inversely proportional to the affinity), and *metadynamics* (insertion of randomly generated new elements).

Immune network algorithm is based on idiotypic network model [6] of N. K. Jerne. Later work from Varela and Coutinho [8] presented *second generation immune networks* that emphasize three characteristics: structure, dynamics, and metadynamics. Discrete models rely on differential equations and iterative procedures, and avoid the problem of finding the concrete analytical solution. De Castro and Von Zuben proposed aiNet (Artificial Immune NETwork) [9] that includes clonal selection, clonal expansion and affinity maturation, as presented in CLONALG, but it introduces term of *clonal memory* and also has different *metadynamics* (elimination of clones with low affinity), *clonal interactions* (defining affinity

between antibody clones), *clonal suppression* (elimination of clones with low affinity to each other), *network construction* (insertion of clones into the network), *network interaction* (defining affinity between antibodies), *network suppression* (elimination of antibodies with low affinity to each other), and *diversity* (insertion of randomly generated new elements).

4. Danger theory

Central idea of Danger theory is context dependant response to invading pathogens, different from traditional self/non-self paradigm, explaining self and non-self coexistence. Immune system cells should be incapable of attacking its host or self, because such cells are eliminated during maturation, but in many cases their behavior is opposite (ignoring injections of non-self proteins, rejection of tumors).

Matzinger's hypothesis [10] provides explanation for above anomalies and states that alarm signal in dying cells stimulates APC and triggers immune system reaction. In that way immune response is not reaction to non-self, but to stimulus or signal dispersed in small *danger area* [3] around the cell. According to Matzinger, there are two types of signal. *Signal one* is binding signal of an immune cell and antigen or an antigen pattern presented by APC. *Signal two* represents alarm signal from T_H cell for B cell activation or co-stimulation signal from APC for T cell activation. This theory resolves T_H cell activation problem by stating that alarm signal activates APC with co-stimulation signal that further activates T cell and than T_H cell stimulates activation of B cell. Basic rule of Danger theory states: *activate if both signals received; die if only signal one received; ignore if only signal two received*. It explains problematic behavior like immunity to self, but is still debated in immunological society. This paradigm presents signaling which leads toward finding a subset of features instead of final set of feature vectors describing non-self.

5. News articles recommendation system

AIS is shown to be a robust and adaptive computational method which we consider suitable for predictions and recommendations in the world of temporal user preferences. Danger signal can be used as an indication of user interest in an article, as suggested in [3]. Our news articles recommendation system consists of AIS algorithm parts from *clonal selection* and *immune network algorithms* combined with a Danger theory model, similar to one employed in AISEC e-mail classification algorithm [11].

In recommender systems latent user preferences are indicated by a wide range of data including user features (user data e.g. age, gender), user behavior with similar preferences (showing interest or rating), and features of the observed elements. News articles, as elements of interest, are characterized by a set of heterogeneous features (e.g. title, authors, date, keywords, etc.) represented as feature vectors. Some of them belong to predefined vocabularies or data structures (author's names, keywords, category), but some are real values (read score, rating) or free text (user opinions, abstract). In this concept, interesting are only features that are easily evaluated, measured, compared and mutated. We are concentrating on keywords and categories combined with user actions. Keywords are contained in word library and associated to the easily mutable feature vector, while categories are presented in a tree structure.

A prerequisite to successful news article recommendation process is that user is authorized and system has already learned some interesting articles features. News articles can be classified as more or less interesting based on the matching (affinity) of their features and user preferences, and good matching can be considered as "signal one". Danger "signal two" is explicitly alarmed when user reads an article for certain time. Additionally, "signal two" can be also raised upon other user actions, e.g. print action, because some users prefer off-line reading of the paper. Reading and printing danger signals are similar according to the representation of user interest. System learns only when both signals (matching and reading) are present. In this way system learns good filtering, used when particular logged user accesses personalized Web portal's news page. Presentation of articles to user is not presented in this paper, but number of different techniques could be used, e.g. category sorted view with descending news sequence (header and first sentence preview). User preferences change in time, so definition of self adapts accordingly.

When user just browses through presented articles without really reading them (reading time below certain duration threshold or lack of read action in certain predefined period), only "signal one" is raised. Features of unread article are of no interest to user and antibodies that matched

this article should be eliminated. The main goal is to produce a set of antibodies that match features which user finds interesting.

Our algorithm works with two populations of B lymphocytes: free B cells in BL set and memory B cells in MBL set representing long lasting memory of good matches. Each memory cell represents a set of interesting article features for a particular user and each antigen represents a new article as shown in comparison and mapping Table 1 of immune system analogies used for AIS+DT recommender system entities.

Table 1. Immune to AIS+DT system mapping

Immune system	AIS+DT system entity
Nonself	interesting article
Antigen	extracted article features
B cell in BL set	set of randomly generated, cloned and mutated features
B cell in MBL set	set of interesting features
Affinity function	feature resemblance

Every article must be processed to the antigen format for comparison with B cells. Fig. 1 presents method new_article(), run on every new article that arrives in the system, performing feature extraction and rating the article according to interesting features set, i.e. highest affinity of antigen and all memory cells.

```
begin new_article(article)
  ag ← process(article)
  article.rate ← 0
  for each mbl of MBL do
    if (affinity(mbl,ag) > ε) then
      if (affinity(mbl,ag) > article.rate) then
        article.rate ← affinity(bl,ag)
end
```

Figure 1: Article rating

All articles introduced to the system are immediately rated, but are not immediately processed for system adaptation and learning, because of predefined period of time in which user can decide if this article is interesting. User action is monitored and appropriate action() method is triggered upon it (Fig. 2).

User feedback acts as a co-stimulation signal. If the user reads (or prints) an article, both signals are present and proliferation and mutation of best free cells are activated. Helper method find_n_best() is used for selecting cells with highest affinity. Cloning and mutation are done via clone_mutate() method which returns a set of mutated clones. If the best mutated clone is better than any existing memory cell, it is promoted to memory cell, and the stimulation level of all existing memory cells with high affinity to this clone is reduced. If the user does not read an article in predefined period or does not print it (lack of signal two) the same method is called with false user_action value and cells with high affinity to that article are eliminated. Estimated value of interest is afterwards asynchronously confirmed or disputed by user action.

```
begin action(ag,user_action)
 if (user_action) then
   for each bl of BL do
     if (affinity(bl,ag) > α) then
       mbl.stim ← mbl.stim + 1
   C ← find_n_best(BL,ag,cFB)
   CM ← clone_mutate(C, affinity(C,a))
   BL← insert(BL,CM)
   bl1 ← find_n_best(BL,ag,1)
   mbl1 ← find_n_best(MBL,ag,1)
   if (affinity(bl1,ag) > affinity(mbl1,ag)) then
     bl1.stim ← cMBLSL
     MBL ← insert(MBL,nmbl)
     for each mbl of MBL do
       if (affinity(mbl,bl1) > β) then
         mbl.stim ← mbl.stim - 1
 else
   for each bl of BL do
     if (affinity(bl,ag) > γ) then
       BL ← remove(BL,bl)
   for each mbl of MBL do
     if (affinity(mbl,ag) > δ) then
       MBL ← remove(MBL,mbl)
 for each bl of BL do
   bl.stim ← bl.stim – 1
   if (bl.stim = 0) then
     BL ← remove(BL,bl)
 BL ← insert(BL,rand(cBLNew))
 for each mbl of MBL do
   if (mbl.stim = 0) then
     MBL ← remove(MBL,mbl)
end
```

Figure 2: Core algorithm

Since cells have finite lifetime, they are eliminated after predefined time if they have not recently recognized any antigen. Cell aging is accomplished through stimulation level downcounters and all cells are set to initial stimulation level during generation process. In each iteration cells with zero stimulation are eliminated, free cells stimulation level is reduced (aging), and new randomly generated cells are introduced as free cells.

Algorithm is tested on 2D real-valued shape space with affinity calculation determined as normalized Euclidean distance (range 0.0-1.0). Initial set of 300 BL and 100 MBL cells was randomly generated. Stable learned set is achieved after presenting approx. 100 interesting articles. Values of threshold parameters α, β, γ, and δ were all set to 0.9. Best 7 BL cells (cFB) were selected for cloning, and initial stimulation levels were set to 25 (cMBLSL) and 15

(cBLSL). Fig. 3 presents MBL population after 10000 iterations with four large clusters defined.

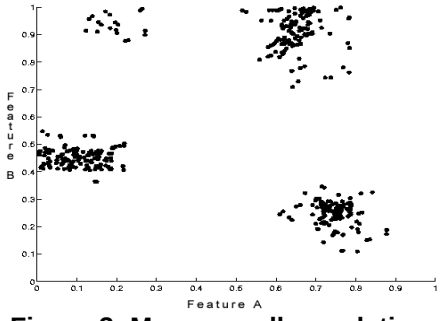

Figure 3: Memory cell population

As presented in Fig. 2, it is obvious that MBL population was stable around 100 cells and that BL cell population varies from 90-550 units.

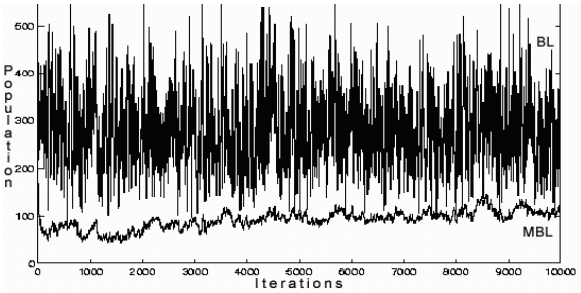

Figure 4: Size of cell populations

6. Conclusion and future work

Recommendation problem does not include finding a single optimum, but rather identification of sorted subset of good choices. Artificial immune system is good approach through antigen-antibody interaction matching mechanism and antibody-antibody interaction diversity mechanism with distributed control. Basically, this is a type of dynamic classification system that is dependant on user's preferences which may change in time. In the learning phase system learns how to rate an article, and in the run-time phase articles are rated and the system adapts and learns accordingly to the user actions.

Future work includes many raised questions about scalability of the described system, because of large amount of data which must be addressed. Comparisons to other continuous learning algorithms are being conducted and first results compared to a variant of naïve Bayesian classifier system show similar results in prediction accuracy. Also, new types of article features which require new mutation techniques will be used. It is also possible to observe the user's ratings of an article or measure user profile similarity with collaborative filtering, well suited for an initial user profile replication.

Principle of context dependant activation and automatic adaptation to changeable user preferences is suitable for recommendation problem and presented algorithm, showing that AIS+DT approach is applicable as a recommender system.

7. References

[1] De Castro NL, Timmis J. Artificial Immune Systems: A New Computational Intelligence Approach. Great Britain: Springer; 2002.

[2] De Castro NL, Von Zuben FJ. Artificial Immune Systems: Part II - A Survey of Applications. DCA-RT 02/00. 1999.

[3] Aickelin U, Cayzer S. The Danger Theory and Its Applications to Artificial Immune Systems. In: Timmis J, Bentley PJ, editors. Proc. of the First International Conference on Artificial Immune Systems: 2002 Sep 9-11; Canterbury, UK. Canterbury: University of Kent Printing Unit; 2002 p. 141-8

[4] De Castro NL, Von Zuben FJ. Artificial Immune Systems: Part I - Basic Theory and Applications. DCA-RT 01/99. 1999.

[5] Burnet FM. The Clonal Selection Theory of Acquired Immunity. Cambridge: Cambridge University Press; 1959.

[6] Jerne NK. Towards a Network Theory of the Immune System. Annals of Immunology 1974; 125C: 373-89.

[7] De Castro NL, Von Zuben FJ. The Clonal Selection Algorithm with Engineering Applications. In: Whitley D, et al, editors. Proc. of the Genetic and Evolutionary Computation Conference 2000 Jul 8-12; Las Vegas, USA. San Mateo, USA: Morgan Kaufmann; 2000. p. 36-7.

[8] Varela FJ, Coutinho A. Second Generation Immune Networks. Immunology Today 1991; 12(5): 159-66

[9] De Castro NL, Von Zuben FJ. aiNet: An Artificial Immune Network for Data Analysis. In: Abbass HA, et al, editors. Data Mining: A Heuristic Approach. Idea Group; 2001. p. 231-59

[10] Matzinger P. The Danger Model in Its Historical Context. Scandinavian Journal of Immunology 2001; 54: 4-9

[11] Secker A, Freitas AA, Timmis J. AISEC: an Artificial Immune System for E-mail Classification. In: Sarker R, et al, editors. Congress on Evolutionary Computation: 2003 Dec 8-12; Canberra, Australia. Piscataway, USA: IEEE Press; 2003. 131-9.