Mimicking immunological memory mechanisms in data analysis

Immunological Algorithms

SŁAWOMIR WIERZCHOŃ
KRZYSZTOF CIESIELSKI
MIECZYSŁAW KŁOPOTEK
Institute of Computer Science, Warsaw
Polish Academy of Sciences
Slawomir.Wierzchon@ipipan.waw.pl
Krzysztof.Ciesielski@ipipan.waw.pl
Mieczyslaw.Klopotek@ipipan.waw.pl

Prof. Sławomir T. Wierzchoń is director of the Department of Artificial Intelligence at the Institute of Computer Science

Dr. Mieczysław A. Kłopotek, an associate professor at the Institute of Computer Science, does research in the field of artificial intelligence

Dr. Krzysztof Ciesielski, an assistant professor at the Institute of Computer Science, studies the grouping of textual data in Internet search engines

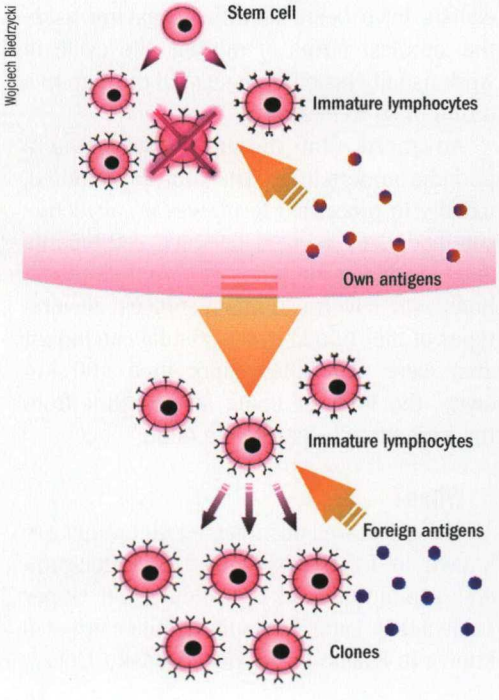
Many innovative computer science techniques mimic nature in some way. For instance, an algorithm inspired by biological mechanisms of immunological memory can be used to identify thematic groups within very large sets of textual documents

Work on developing artificial immune systems (AIS), underway for nearly two decades now, represents one avenue of a broader domain of research on biologically-inspired computing methods, embracing neural networks, genetic algorithms (inspired by the mechanisms of evolution), and swarm algorithms (based on the collective, self-organized behavior of individual units, such as ants or birds). Artificial immune systems harness algorithms inspired by the mechanisms identified by theoretical immunology. But how can the biological mechanisms of the immune system be used for searching data? To understand that, first let's take a closer look at how the immune system works.

System of defense

The main actors involved in the immune system's defensive reaction are lymphocytes, a kind of white blood cell. They are classified (based on their place of origin) into two basic groups, known as B-cells and T-cells. The surface of each B-cell has certain receptors called antibodies, which are proteins capable of binding antigens (i.e. foreign bodies like bacteria, viruses, fungi, etc.) that pose a threat to the organism. The characteristic part of each antigen which gets recognized and bound by an antibody is called an epitope or antigenic determinant; each basic determinant type is called an idiotype. Similarly, the particular fragment of an antiboty which actively binds to the epitope of a given antigen is called a paratope. While antigens possess only epitopes, antibodies have both epitopes and paratopes.

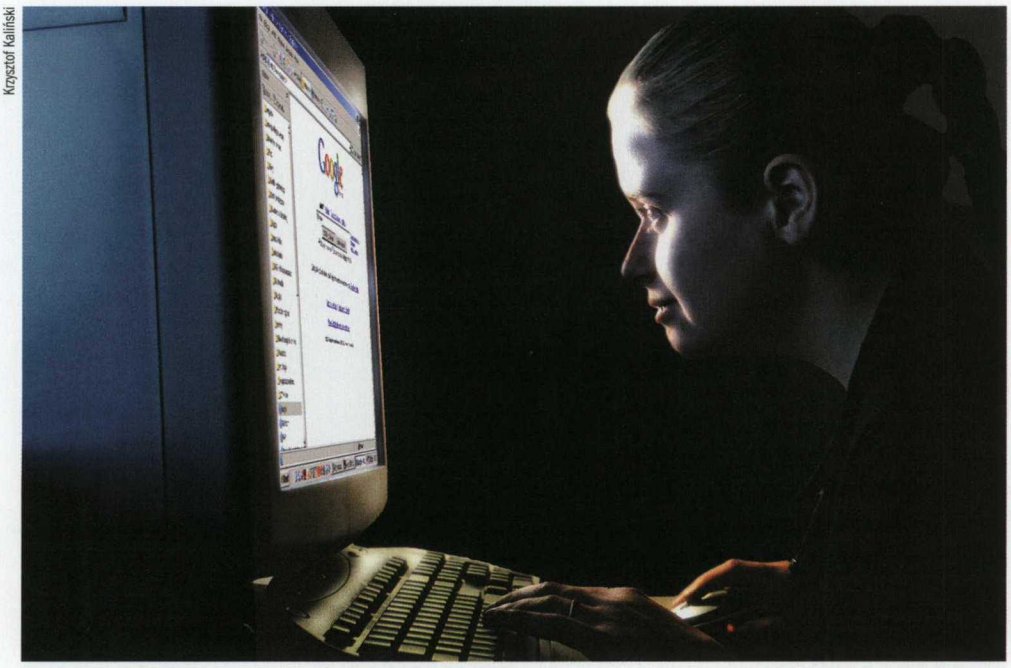
Real paratopes and epitopes are 3D structures. If they are complementary in terms of their geometric, physical, and chemical properties, we say that a paratope recognizes or binds a presented epitope. To study the interactions between epitopes and paratopes, the concept of a "shape space" was introduced, i.e. a multidimensional space whose various dimensions correspond to specific characteristics of the molecules analyzed. In this context, the specificity of the epitope-paratope bond can be treated as the degree of similarity between the two molecules, most often defined as



Wojciech Biedrzycki

The mechanism involved in the clonal selection process, optimizing the body's immune response. In the presence of a dangerous antigen, only those lymphocytes which possess receptors recognizing that antigen will become activated. The number of lymphocytes which bind to the right foreign antigens sharply increases and that quickly leads to the production of specific resistance to the pathogen

Krzysztof Kaliński



Huge collections of textual data are commonly encountered by everyone these days, e.g. when utilizing Internet search engines. To enable vast sets of data to be searched quickly, more efficient algorithms are constantly being devised

a function inversely proportional to the distance between the points representing those molecules within the shape space.

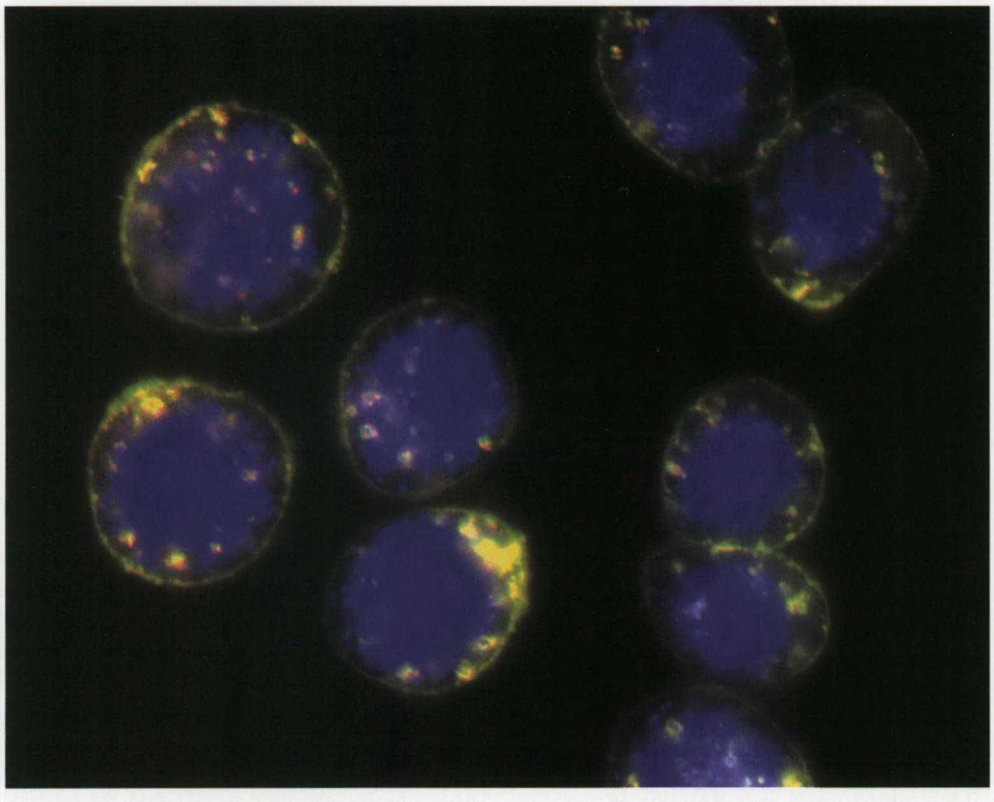
If a sufficiently large number of instances of a given antigen are introduced into an organism which has not previously had contact with it, they give rise to the primary immune response, involving what is called clone expansion and somatic hypermutation. The first of these terms refers to the rapid copying or cloning of those B-cells whose antibodies bind most strongly to the presented antigens. To make the resulting clones more effective at fighting the antigens, they are subject to a very intense process of mutation. The mutated effective cells release antibodies into the organism's fluids, where they may flow throughout the organism and eliminate the threat. Effective B-cells are subject to further cloning and somatic hypermutation. At the same time, cells not participating in the immunological response get eliminated from the organism. This process continues until the antigen concentration drops below a certain threshold. The mechanism described here is called clonal selection.

The immune system response evidences a certain tolerance: if the concentration of epitopes is very low or very high, the organism will not react to the presented antigen. Only an "average" antigen dose triggers a defensive reaction.

Antibody paratopes are stimulated by all the epitopes present in the organism, regardless of whether they are from antigens or antibodies. That is because a new antibody, let's call it Ab_1, represents a new protein present in the body, and its production during the clonal expansion process provokes a new response from the organism leading to the production of a new type of antibody, let's call it Ab_2. In general, the production of antibody Ab_i provokes the production of successive types of antibodies, the successive generations of proteins forming what is known as an idiotypic chain. Such a chain is characterized by self-sustainability, meaning it can exist even when the epitope which initiated its development is completely eliminated from the organism. The renewed appearance of the initial antigen then triggers the nearly immediate production of effective antibodies. This self-sustainable idiotypic chain can be treated as a model of "immunological memory," and this phenomenon whereby "remembered" antibodies are produced to effectively fight a given type of epitope is called the secondary immune response.

Exploratory data analysis

One important algorithm inspired by both the theory of clonal selection and the theory of idiotypic chains is called aiNet, modeling not B-cells but antibodies. Both



Invitrogen Molecular Probes, www.probes.invitrogen.com

Although they do not remind one of computers, human immune cells actually go about performing their tasks in the body based on effective algorithms

the epitopes of antigens and the paratopes of antibodies are here treated as points in an n-dimensional Euclidean space. The antibody generation process consists of two main stages. The first utilizes the principle of clonal selection and affinity maturation. In the second stage, interactions occur between the cells of the system and cells differentiate in keeping with the theory of idiotypic chains. In specific, the similarity between cells is noted and if it exceeds a set threshold, the overly similar cells are eliminated in order to reduce redundancy within the resulting immunological memory. Moreover, overly specialized cells, i.e. those antibodies which recognize only a small number of antigens, are also removed from memory. Each reduction phase is followed by the introduction of new (randomly generated) cells, to ensure the diversity of the immunological repertoire as modified in successive cycles.

The solution described here exhibits a certain kind of data compression. Antibodies, which can be treated as prototypes, are situated in strategic areas occupied by antigens. Possessing such a reduced set of points reflecting the most essential traits of the data set (antigens), one can proceed with their analysis. It is important that such analysis is not performed on the original set of antigens, but on the smaller set of antibodies. Moreover, this algorithm does not require a fixed number of classes which the data is to be classified into, and is thus excellently suited to incremental learning. Adding new data does not require reanalysis of the aggregated data content (as is the case for classical cluster analysis algorithms), but only entails potential modifications to the set of memory cells.

Analysis of large document sets

Our team has proposed a method for analyzing and representing the content of large sets of textual data which involves harnessing the advantages of the immunological algorithm as described above for the purposes of contextual document analysis.

Like most information retrieval and processing systems, our approach represents documents in the form of vectors $v_d = (v_{d,1}, ..., v_{d,T})$, where the subscript d refers to a document, T is the number of terms (expressions), and $v_{d,t}$ is the weight – the product of the number of occurrences of term t within the given document, and the logarithm of the overall number of docu-

ments divided by the number of documents which contain the term *t*. Such defined weight is called the *tfidf* weight (term frequency - inverse document frequency).

The novelty of our solution consists in its contextual approach, replacing a uniform scheme for evaluating the relevance of words and phrases within the whole collection (expressed by the tfidf weight) with an evaluation that accounts for local distributions of the occurrence of individual terms within groups of documents of similar topic. Each set of documents represented using a local schema of term weights forms a contextual group. This approach is independent of the chosen grouping model. The specific criteria that are optimized during the identification of context groups include balancing the numerical size of the individual groups, the homogeneity of topic weight distribution within a group (topical uniformity), and establishing a division in the term space (dictionary) in tandem with a division in the set of documents.

Relations between distinguished contexts and the documents belonging to them are presented in the form of document maps, an approach which has two advantages. On the one hand the creation of such a map allows the user to gain a quick overview of the content of document set D, and on the other it distinguishes groups (not necessarily disjoint ones) of "similar" documents within the set D. Operating with such groups greatly speeds up the process of seeking interesting documents. Map creation is therefore a process of detecting the internal structure of a set of objects, plus the visualization of the structure so detected.

The effectiveness of this new approach has been confirmed by experiments on a set of 20,000 user postings on 20 different discussion groups. Two alternative methods were tested for initializing the initial immunological memory (i.e. the set of antibodies, each of which characterizes a certain basic topic of discussion). The first method, called contextual initialization, involved a conscious choice of a certain number (smaller than the number of discussion groups) of antibodies to describe clusters containing documents of similar topic. The other method involved the random initialization of the vectors representing the antibodies.

We found that the proper initialization of the opening set of antibodies, combined with our contextual approach, has a significant impact on the time taken to identify idiotypic chains: the context model required some 10 minutes, while the standard aiNet method required more than 20 hours of learning. One of the reasons for this is that antibodies correctly chosen in the initial stage of learning are capable of long-term survival within the immunological memory. That finding means that the level of antibody maturity partially depends on the correct initialization of memory and affects the convergence of the whole algorithm.

The observed results indicate that immunological algorithms, which in essence represent a fusion of swarm and genetic algorithm techniques, could become a universal tool for resolving problems of various degrees of complexity. ■

Further reading:

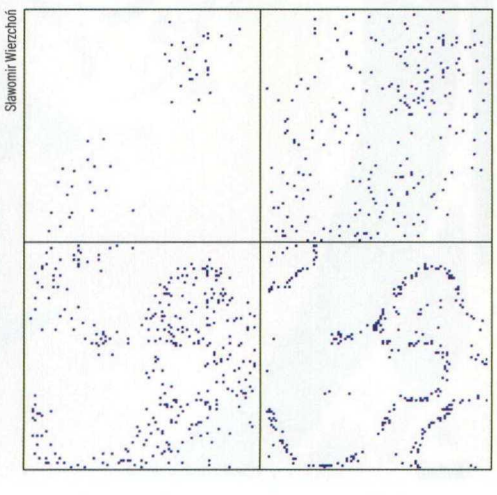
Ciesielski K., Wierzchoń S.T., Kłopotek M.A. (2006). An immune network for contextual text data clustering. *Proc. of the International Conference on Artificial Immune Systems*, 432-445.

Ciesielski K., Kłopotek M.A., Wierzchoń S.T. (2008). Term distribution-based initialization of fuzzy text clustering. *Proc. of the International Symposium on Methodologies for Intelligent Systems.*

De Castro L.N., Timmis J. (2002). *Artificial Immune Systems: A New Computational Intelligence Approach*. London: Springer-Verlag.

Wierzchoń S.T. (2007). *Zastosowanie algorytmów immunologicznych w eksploracyjnej analizie danych.* [*Use of Immune Algorithms in Exploratory Data Analysis*]. [In:] Kulczycki P., Hryniewicz O., Kacprzyk J. *Techniki Informacyjne w Badaniach Systemowych*. Warsaw: WNT.

http://www.ipipan.eu/~klopotek/BEATCA



Sławomir Wierzchoń

The successive stages of forming a self-sustaining structure (from upper left to lower right): after 1300, 1500, 2000, and 4000 iterations, the final structure is nearly unchanging