An Artificial Immune System Approach To Semantic Document Classification[1]

Julie Greensmith, Steve Cayzer
Digital Media Systems Laboratory
HP Laboratories Bristol

HPL-2003-141
July 16th , 2003*

E-mail: Julie.Greensmith@hp.com, Steve.Cayzer@hp.com

artificial immune system, semantic web, document classification, machine learning

AIRS, a resource limited artificial immune classifier system, has performed well on elementary classification tasks. This paper proposes the use of this system for the more complex task of hierarchical, multi-class document classification. This information can then be applied to the realm of taxonomy mapping, an active research area with far reaching implications. Our motivation comes from the use of a personal semantic structure for ease of navigation within a set of Internet based documents.

* Internal Accession Date Only Approved for External Publication

[1] 2nd International Conference on Artifical Immune Systems, Edinburgh, U.K. September, 2003

An Artificial Immune System Approach To Semantic Document Classification

Julie Greensmith & Steve Cayzer

[1]Hewlett-Packard Labs, Stoke Gifford,
Bristol, UK
BS34, 8QZ
Julie.Greensmith@hp.com, Steve.Cayzer@hp.com
http://www.hpl.hp.com/research/bicas/

Abstract. AIRS, a resource limited artificial immune classifier system, has performed well on elementary classification tasks. This paper proposes the use of this system for the more complex task of hierarchical, multi-class document classification. This information can then be applied to the realm of taxonomy mapping, an active research area with far reaching implications. Our motivation comes from the use of a personal semantic structure for ease of navigation within a set of Internet based documents.

1 Introduction

The explosion of information available on the Internet makes it increasingly difficult to navigate. The notion that intelligent annotations to Internet pages should be added is one possible way to approach this problem. Indeed, the underlying idea behind the Semantic Web is to introduce “an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation”[2]. The classification of such Internet documents into taxonomies is one method of enabling the end user to derive information more effectively through the use of defined ‘semantics’. However, such manual annotation of pages is time consuming and for it to work successfully would require the majority of pages to contain such annotation. This would not only mean extra work during the creation of the pages, but would also have to be applied to many existing documents, and many amateur web authors would not have the time, or indeed the patience, to perform such a chore.

One solution which alleviates the problem involves applying a taxonomic structure to web documents in a post-hoc manner, independently of their authors, making searching for relevant content on the Internet a less daunting and more efficient task. As with many other problems involving classification or decision-making in a complex, distributed system, solutions have been inspired by a variety of biological metaphors. One idea in particular appeals as a potential candidate to perform this task, namely *artificial immune systems.* This is inspired by the ability of the human immune system to effectively distinguish between proteins, hence acting as a natural classification system. The modelling of such a system for a functionally similar task is a seemingly logical step.

Indeed, artificial immune systems (AIS) have demonstrable potential for success within this domain, since powerful classifier systems have been developed and evaluated as described in previous work from Watkins & Boggess [24] and Twycross & Cayzer [23]. This paper outlines the use of such a system within the field of Internet document classification. The reader will be introduced to the variety of methods and tools associated with the classification task and the use of taxonomies, in addition to the use of an AIS in the novel area of document classification. The paper is organised as follows: Section 2 introduces relevant information regarding artificial immune systems, feature extraction/representation, classifier systems and taxonomies; Section 3 discusses the work already performed and the research that will be carried out; the final sections include a summary and references.

2 Context

This section introduces the important concepts employed in this area of research. Basic immune system principles are summarised, with particular attention paid to the components that are specifically relevant to our system. The relevant algorithms and mechanisms are explained and related to the task of document classification. Following this, the techniques involved in document classification are examined including feature representation, similarity metrics and the utilisation of a taxonomic structure within this context.

2.1 Immune Systems

At a high level of abstraction, the immune system can be subdivided into two components: *innate* and *acquired.* The acquired component is of particular interest due to the adaptive nature of some of its constituents, that are thought to perform a biological classification task. B-Lymphocytes or B-Cells are one such constituent and within the immune system are responsible for the production of *antibodies* and the development of *memory cells.* An overview of the immune system can be found in many immunology texts, for example Janeway *et al.*[10]

Antibodies are proteins that are produced from B-Cells, in response to the detection of a foreign protein or antigen. The antibodies produced from B-Cells can perform complementary pattern matching with a corresponding antigen, initiating a series of events, resulting in the destruction of the invading pathogen. Each segment of antibody is encoded by a widely separated gene library which is subject to alternate splicing, therefore a wide variety of subtly different proteins can be produced. The pattern matching ability of a B-Cell antibody is *resource limited*. During the maturation period, if an antibody does not successfully match any given antigen proteins, then the resources for the replication of that antibody are removed (Fig. 1). It is thought that the energy used in the complementary antigen-antibody binding is a measure of an *affinity threshold*. Once a B-Cell has exceeded this threshold, the cell is cloned and daughter cells are produced as a slightly mutated version (*hypermutation*). This can improve the antibody-antigen binding as the newly formed antibody may have a higher antigen affinity than the original cell.

Memory cells, formed as a result of this clonal selection process, are B-Cells that are successful with respect to antigen matching. These cells are longer-lived than normal B-Cells and are thus available (perhaps many years later) when a similar antigen challenge is encountered. Under that circumstance, a rapid secondary response can be initiated.

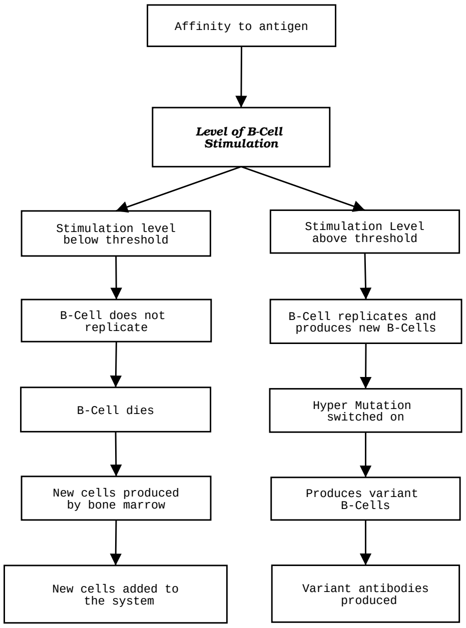
Affinity to antigen
Level of B-Cell Stimulation
Stimulation level below threshold
Stimulation Level above threshold
B-Cell does not replicate
B-Cell replicates and produces new B-Cells
B-Cell dies
Hyper Mutation switched on
New cells produced by bone marrow
Produces variant B-Cells
New cells added to the system
Variant antibodies produced

Fig. 1. The resource limited process of B-Cell regulation [10]

In addition to the action of B-Cells, the adaptive immune response is also facilitated through the use of T-Cells, which aid in the destruction of pathogens and the generation of memory cells. However, the action of T-Cells is not used in our system and so is not explained in this paper.

2.2 AIRS - A Resource Limited Immune Classifier System

Artificial immune systems do not attempt to be an accurate representation of the immune system, but use relevant mechanisms of action and concepts in order to create a robust, decentralised, adaptive system which are used in a variety of classification tasks. In the case of the system we will focus on, AIRS, inspiration has been derived from B-Cells.

AIRS is a particular type of artificial immune system, specifically designed for supervised classification tasks [24]. The system contains a population of B-Cells which respond to virtual antigens derived from the input data. A concentration of identical B-Cells is represented as an *artificial recognition ball* (ARB). Initial populations of both ARBs and memory cells are created from the training data. In order to achieve successful classification, training data must be presented to the system. Each item of training data is encoded as a single antigen, and AIRS responds to the antigen based on the closest matching existing memory cell, which gives rise to the creation of ARBs. The ARBs are then allocated resources based on the success achieved with respect to the antigen matching, i.e. proportional to antigen affinity. At this point, ARBs that are unsuccessful are removed from the system. This value for affinity is used to determine the number of clones produced by that particular ARB. This process is repeated until the average affinity for all of the ARBs is above a pre-determined stimulation threshold. The best matching ARB for a class of antigen is then promoted to a become a *candidate memory cell.* If the affinity of the candidate memory cell is greater than that of the closest memory cell, then the candidate memory cell is placed in the 'pool' of memory cells used in future classifications. This process has similarities with other clustering methods, e.g. k-means [7].

Once the system has been trained in this manner, it is applied to a set of test data. In this, the created memory cells are used as *class prototypes.* A k-nearest neighbour approach is used in order to determine the best match between memory cell and test antigen. The position of the antigen within the space is represented and compared with surrounding memory cells. A radius is established which encompasses a specified amount of neighbours, 'k'. Within this radius, the class of antigen is given by the class of the majority of nearest neighbours .The scenario of a tie is dealt with in various different ways as discussed in [11].

In order to use the AIRS system in the domain of document classification, the system needs no major modification. The system will already work with a variety of feature representations and similarity metrics. Some modification of the configuration file is likely to be required to tune the system for the current task. This includes the correct amount of training data needed to seed the program initially, and the thresholds for resource allocation and mutation rate of the artificial memory cells. The most significant task however involves extracting the information from the documents and placing the information in the correct representation for the system to use.

2.3 Text Classification

Text classification is a relatively well explored area (for an overview see [1]) with a number of potential approaches. In this section, we outline the relevant details of our text classification task, and how we intend to tackle semantic classification. Note that we have not (yet) comitted to any single method; rather, we outline the overall issues in this section and describe some candidate solutions.

Successful classification of the documents relies on attributes of the document, also known as *features*, to be extracted. The weighting and selection of such features has to be deduced, and represented in a way that can be used by the classifier system. This is the process of compilation of *feature vectors*. Features can be *textual*, in the form of extracted words, or *semantic*: for example, a position within a meaningful structure such as a taxonomy.

AIRS also requires the definition of a *similarity measure*, that is, a metric that tells us how 'close' a document is to a memory cell, using the features defined above. Once the representation scheme and similarity metric are fixed, AIRS can be used essentially without modification to perform semantic classification.

Classification using Textual Features: The extraction of text based features is a relatively trivial task, and simply involves extracting the words from the document. The selection of the important features after this extraction has been performed is more complicated. This stage is necessary in order to reduce the dimensionality of the feature vector, so the classifier is not confused by irrelevant terms and to reduce the computational expense. Methods used in feature selection include information gain [20] which is a statistical entropy based measure, and singular value decomposition [3] which is a matrix compression algorithm. In this instance, information gain will be used in order to determine the important words derived in a class of documents (based on a training set of pages). From this information, the 'gain' derived from each word with respect to the usefulness of the word in the classification of the document can be used as a metric. The presence or absence of the words within a class can be translated into a feature vector, and can then be represented in a variety of different ways. However information gain requires the dataset to be static as it depends upon the presence of all the words contained within the entire document collection.

Once the important features have been selected, then the data must be represented in a way that is suitable for use by the classifier system. There are a number of methods that can be implemented in order to provide this representation. Boolean vectors are utilised in Twycross & Cayzer [23] where a 0 or a 1 is used to represent the absence or presence of an important term within a feature vector, and is an effective way of representing the information to the system. Alternatively, term-frequency, inverse document frequency *(tfidf)* vectors can be used. This method takes into account the amount of times a term occurs in a document, and the amount of times that word appears more than once within a collection of documents. The latter is inverted to give a high value for an important word that occurs several times within a document, but is only present within the minority of documents [22]. The resulting vectors have to be normalised to achieve similarity of size. We will investigate both binary and vectorial features in our classification work. Note that we are concerned here with static classification tasks. A possible future extension would involve the use of dynamic document collections, for which alternate representations, capable of incremental update [15] would be more suitable.

AIRS' default choice of similarity measure is Euclidean, which we intend to use as a baseline. The problem with this method is that it regards all attributes as equally important. In reality, the importance of an attribute is likely to depend both on the attribute in question and on the region of feature space in which the memory cell resides. Therefore, we expect the task to benefit from more refined similarity measures, such as those involving wildcards [23], variable thresholds [17] or even variable weights for each attribute.

Classification using Semantic Features: Taxonomies describe a structured way in which various information and classifications can be viewed, with each branch of the taxonomic tree having some relationship with the parent and siblings. They are a simple version of an ontology, which has been described as a "formal specification of a shared conceptualisation" [8]. That is, ontologies are concerned with methods for knowledge representation, and taxonomies are one such method. Taxonomic structures can in the first instance provide a useful classification of documents which are placed into the appropriate taxonomic location. However, it would be useful to classify a document according to multiple, user defined taxonomies. This would provide the user with a *personalised semantic structure* which would be both more comfortable and intuitive. This area of research has been prompted by the need to examine the classification of document within a hierarchical structure without the reliance of rigid logic techniques, to provide a system that has more flexible and ultimately more meaningful navigation experience. In order to achieve a classification based on this multiple mapping it must also be appreciated that the positioning of a document in one taxonomy can also be a feature of that document i.e. a *semantic feature.*

The representation of such a feature is fairly simple - we can use string descriptors such as '/computers/languages/java' and '1.3.6', or alternatively an index into an externally referenced taxonomic scheme. The more complex issue is that of computing taxonomic similarity. Here we have a range of approaches to choose from, for example a 'probability of the least subsumer' measure [19] or an upwards cotopy [13]. A discussion of the issues surrounding taxonomic similarity measures within the context of AIS can be found in [14].

2.4 Related Work

The classifier system used in this particular instance relies on the adaptive nature of the immune system and its ability to remember encountered antigens in a content addressable manner. However, there are many techniques available for classification. Such methods include the Naive Bayesian classifiers, nearest neighbour, decision trees and neural networks. Naive Bayesian classifiers are a common example of a probabilistic classifier, that is that the probability that a vector represented collection of weighted terms from a document belongs to a class of document types [22]. The k-nearest neighbours approach (as used by AIRS) relies on the assumption that in the feature space, if a document x is closest to k documents, then it belongs to the class of the majority of those documents [7]. The application of a neural net to perform classification is less common than the use of Naive Bayesian classifiers, but can be useful, especially where weighted terms are involved.

Classification of web documents has already been performed by a series of different research groups. Indeed, there is a web toolkit, Rainbow [12] available, which allows the user to experiment with Naive Bayes, k-nearest neighbour, tfidf, and probabilistic indexing techniques. A more recent and relevant example is the work of Twycross & Cayzer [23] who used a co-evolutionary AIS for the classification of web documents. We are extending the AIS beyond 2-class problems to classification tasks that are multiclass and hierarchical. Ceci et al [5] compare a Naive Bayes and centroid based technique for just such a hierarchical document classification task, finding that the Naive Bayes approach is generally better. This is interesting, because there is evidence that an immune approach can outperform Naive Bayes for document classification tasks [23]. In fact, we are testing AIRS on a similar corpus to that used by Ceci, but we also plan to perform an ontology transformation using semantic features.

3 Using AIRS As A Semantic Classifier

3.1 Validation of AIRS

The evaluation of the AIRS system has to date formed the major component of the work performed toward this project. This validation is performed in order to ensure that the system can perform supervised classification based upon known test database information. The ability of an artificial immune system to achieve the necessary classification task is derived from the fact that the system relies on the local interactions between individual antigens and antibodies. This in turn makes the system robust and gives it the ability to handle situations where traditional techniques might not be able to cope. Previous tests using this system [11] indicate the performance in classification of multiple class data to be exemplary. The ability of the system to remember previous actions in a content addressable manner, irrespective of centralised control makes it an ideal learning paradigm [9].

The initial testing and validation of the system used a voting dataset obtained from the UCI repository [4]. This dataset is an attractive one to use for many reasons. Firstly, this is a popular dataset, providing us with many figures against which we can compare the performance of AIRS. Secondly, a co-evolutionary AIS system [23] has already provided exemplary performance on this dataset. We have chosen AIRS rather than the co-evolutionary model for our document classification task due to the proven ability of the former on multi-class data. A comparable performance on this simpler task will give us confidence that we are not choosing an inferior classifier architecture. Thirdly, AIRS has not yet been tested on the voting dataset, and so the results will be interesting on their own merit.

The voting dataset contains votes from the 1984 US Congress on 16 different issues, and from the voting pattern, the senators are classified as democrat or republican accordingly. The information was extracted from this source in a 16-D vector, with equal weighting to each vote. The representation of the features was based on a system of a score of 0 for a 'no' and 1 for 'yes', and 2 for 'abstain'. As there are an equal number of features for every potential antigen, there was no need to normalise the data. In order to complete the process, 10% of the voting dataset was excluded from the dataset and used as the test data, and was repeated several times with different parts of the dataset comprising the test file. This gave a cross-validation of the system. On successful compilation of the AIRS code (using g++ version 2.94), the classifier was run and various statistical outputs were collected. Although a full analysis has not been performed, initial indications are promising, showing a classification accuracy in excess of 95%, which compares well with tests performed using a Naive Bayesian classification system (90%) and the co-evolutionary artificial immune system (97%) developed by Twycross & Cayzer [23]. Note that this result was obtained using a naive feature representation with a Euclidean similarity metric. Intuitively, one would expect to improve performance through the use of techniques like wildcards. Nevertheless, it is encouraging that AIRS appears to offer competent performance levels using only its default configuration.

3.2 Taxonomic Classification

Once the validation results have been analysed, the focus of the project will turn to the classification of actual web documents. This will involve the extraction of the information from the documents into feature vectors, the selection of appropriate features and the representation using feature vectors. The preparation of the feature vectors is a multistage process, including extraction of all of the words, followed by parsing of the documents to remove 'fluff words' and HTML tags. In order to further reduce the dimensionality of the document, a Porter Stemming Algorithm [18] is to be implemented which reduced the words in the document to their stem, e.g. 'classifier' and 'classification' would be reduced to 'classi'. Further dimensionality reduction will be achieved through the use of an information gain algorithm which will provide a measure of the k-most informative words within a class of document. The absence or presence of a word within a document can be represented as either a boolean value or as some frequency dependent vector (which would be normalised), both of which will be in suitable format for the use of the AIRS system. From this point the documents will be classified into an established taxonomic structure e.g. provided by Yahoo!. This is not only a useful task in its own right, it also provides an interesting test of the AIRS system's extensibility. Not only is the classification task concerned with documents (hence the textual features), it also requires multi-class output. The classification structure achieved is hierarchical which is a novel domain for the use of AIRS.

3.3 AIRS as a Tool for Personalised Semantic Structure

The subjectivity of the placing of the documents within this structure is a serious issue, as classification schemas, derived by human or machine annotation, can be inconsistent and often there is no 'correct' answer. This fact is not ignored and consequently, the process is adapted to include the additional option of multiple taxonomy mapping. In general, work performed regarding ontology mapping is on a one-to-one mapping basis, as demonstrated in [6] and reviewed by [21]. This implies that the positioning of a document in one taxonomy determines the placing of a document in an alternative taxonomy. However, we think that it is more effective to use *both* the features extracted from the document *in addition to* the position in an initial taxonomy in order to create the personalised semantic structure.

We start with a set of documents that have been pre-classified into a taxonomic structure. To derive the placing of the documents in an alternative taxonomy, a selection of documents from this original taxonomy are used as training data. The features of the training items are derived from the textual features of the documents, and the semantic feature (position) within the original taxonomy. Once the training has been performed with this information, AIRS classifies test documents using textual features and the semantic feature of the original taxonomy, into the new taxonomy. The success of this technique would have implications in a number of other parallel problems, such as data integration and machine translation.

We hope to use AIRS as a more effective searching and navigational tool for use on the Internet, through using a robust and adaptable technique, which gives the user the choice on the semantic structure of the taxonomy of the web documents. Additionally the application of this system in a context where the collection of pages is constantly changing would be an interesting test of the extensibility of the system as a whole. Indeed, there is evidence that AIRS is well suited for a dynamic environment [16]. The internet is, of course, inherently dynamic, and so extending document classification in this way is an important issue. The interesting question for us is not how can the AIRS work as a classifier, but rather how far can we push the system into performing increasingly complex tasks. As quoted in Marwah & Boggess [11], "..since AIRS is a very recent classifier, there are many possible areas for the exploration of the algorithm", and this seems like an excellent opportunity to do just that.

4 Summary

As the amount of information on the Internet increases, more meaningful and powerful navigational tools must be developed. An artificial immune system that has previously performed well in other classification tasks, will be used to classify web documents into a taxonomic structure. This structure will necessarily be subjective, therefore the information can be mapped to alternative taxonomies in order to provide the user with a personalised semantic structure. This not only attempts to introduce a bio-inspired technique to the development of the semantic web, but will test some of the boundaries involved in using an artificial immune system in the domain of document classification.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press, New York, 1999.
2. Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, May 2001.
3. Daniel Billsus and Michael J. Pazzani. Learning collaborative information filters. In *Proc. 15th International Conf. on Machine Learning*, pages 46–54. Morgan Kaufmann, San Francisco, CA, 1998.
4. C.L. Blake and C.J. Merz. UCI repository of machine learning databases (available at http://www.ics.uci.edu/~mlearn/mlrepository.html), 1998.
5. Michelangelo Ceci, Floriana Esposito, Michele Lapi, and Donata Malerba. Automated classification of web documents into a web hierarchy. In *Proceedings of the international IIS: IIPWM2003*, pages 59–68, Zakopane, Poland, June 2003. Springer.
6. AnHai Doan, Pedro Domingos, and Alon Y. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *SIGMOD Conference*, 2001.
7. AA Freitas. *Data Mining and Knowledge Discovery with Evolutionary Algorithms.* Spinger-Verlag, Berlin, 2002. Page 34.
8. T. R. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands, 1993. Kluwer Academic Publishers.
9. John E. Hunt and Denise E. Cooke. Learning using an artificial immune system. *Journal of Network and Computer Applications*, 19(2):189–212, 1996.
10. Charles A. Janeway, Paul Travers, Mark Walport, and Mark Shlomchik. *Immunobiology The Immune System in Health and Disease.* New York : Garland ; Edinburgh : Churchill Livingstone., 2001.

11. Gaurav Marwah and Lois Boggess. Artificial immune systems for classification: Some issues. In *Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS)*, pages 149–153, Canterbury, UK, September 2002.
12. Andrew McCallum. Rainbow (available at http://www-2.cs.cmu.edu/ mccallum/bow/rainbow/), 1998.
13. A. Mdche and S. Staab. Measuring similarity between ontologies. In *Proc. Of the European Conference on Knowledge Acquisition and Management - EKAW-2002*, Madrid, Spain, October 1-4, October 2002. Springer.
14. T Morrison and U Aickelin. An artificial immune system as a recommender for web sites. In *Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS)*, pages 161–169, Canterbury, UK, September 2002.
15. Nikolaos Nanas. An adaptive, evolutionary user profile for knowledge management. KMI-TR-114, http://kmi.open.ac.uk/publications/techreports-text.cfm.
16. M Neal. An artificial immune system for continuous analysis of time-varying data. In *Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS)*, volume 1, pages 76–85, Canterbury, UK, September 2002.
17. Nasraoui O., Gonzalez F., and Dasgupta D. The fuzzy artificial immune system: Motivations, basic concepts, and application to clustering and web profiling. In *IEEE International Conf. on Fuzzy Systems (IEEE-FUZZY 2002)*, pages 711–716, May 2002. Part of the World Congress on Computational Intelligence (WCCI) held in Honolulu, HI, USA, May 12-17, 2002.
18. M.F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980. reprinted in Sparck Jones, Karen, and Peter Willet, 1997, Readings in Information Retrieval, San Francisco: Morgan Kaufmann, ISBN 1-55860-454-4.
19. P.W.Lord, R.D. Stevens, A. Brass, and C.A.Goble. Semantic similarity measures as tools for exploring the Gene Ontology. In *Pacific Symposium on Biocomputing*, pages 601–612, 2003.
20. J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
21. Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal: Very Large Data Bases*, 10(4):334–350, 2001.
22. Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47, 2002.
23. Jamie Twycross and Steve Cayzer. An immune-based approach to document classification. In *Proceedings of the international IIS: IIPWM2003*, pages 33–48, Zakopane, Poland, June 2003. Springer.
24. Andrew B. Watkins and Lois C. Boggess. A Resource Limited Artificial Immune Classifier. In *Proceedings of Congress on Evolutionary Computation*, pages 926–931. IEEE, May 2002. Part of the 2002 IEEE World Congress on Computational Intelligence held in Honolulu, HI, USA, May 12-17, 2002.