An Artificial Immune System Approach to Document Clustering

Na Tang
Computer Science Department
University of California, Davis
CA 95616, USA
natang@ucdavis.edu

V. Rao Vemuri
Computer Science Department
University of California, Davis
CA 95616, USA
rvemuri@ucdavis.edu

ABSTRACT

It has recently been shown that artificial immune systems (AIS) can be successfully used in many machine learning tasks. The aiNet, one such AIS algorithm exploiting the biologically-inspired features of the immune system, performs well on elementary clustering tasks. This paper proposes the use of the aiNet to more complex tasks of document clustering. Based on the immune network and affinity maturation principles, the aiNet performs an evolutionary process on the raw data, which removes data redundancy and retrieves good clustering results. Also, Principal Component Analysis is integrated into this method to reduce the time complexity. The results are compared with some classical document clustering methods - Hierachical Agglomerative Clustering and K-means.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering

Keywords

Artificial immune system, document clustering

1. INTRODUCTION

Document clustering, i.e., unsupervised document categorization, is a very important and challenging problem in the area of information retrieval and text mining. It has been proposed for use in navigating and browsing document collections [6] or as a tool for Web search engines [8, 12]. The Hierarchical Agglomerative Clustering (HAC) and K-means are two commonly used clustering techniques for document clustering [9]. HAC starts with all data points each in its own cluster, and repeatedly merges two closest clusters into one cluster. It finally generates a hierarchical grouping of data. The K-means aims to find K clusters by starting with K randomly selected centroids and then repeatedly assigning all points to the closest centroid and re-computing the centroid of each cluster. Besides HAC and K-means, many other methods for document clustering also exist in the literature. Some of them are: improved versions of K-means such as bisecting K-means [9] and incremental K-means [11], Suffix Tree Clustering [12], etc.

Most of these methods directly apply clustering techniques to the raw collection of documents. However, with the explosion of information available electronically, the size of document collections is becoming increasingly large. In the case of large collections, more noise exists in the data, which causes inferior clusters. In this paper, a method is proposed to perform a preprocessing before the clustering procedure. The preprocessing reduces the noise and redundancy information and retrieves more compact clusters, and eventually improves the quality of the clusters. Due to this reason, this step is called refining preprocessing.

The proposed document clustering approach employs the aiNet (Artificial Immune Network) [4, 3], an immune system based algorithm which combines the desired preprocessing and clustering procedures. The entire framework is shown in Figure 1. The "representation and feature selection" phase is handled by converting all documents into a matrix comprised of multidimensional vectors. The aiNet first builds a compressed data representation for the vectors through the refining preprocessing (compressing the rows) and then automatically detects clusters from the compressed data via clustering techniques. The original aiNet algorithm uses HAC to detect clusters. In this paper, both HAC and K-means are used. Finally the clusters of documents are recognized based on the relation between the compressed data and the original document set. Principal Component Analysis (PCA) is introduced as an option to reduce the dimension of the vectors (compressing the columns), which in turn results in a speedup the compression process and further reduces the noise from the data.

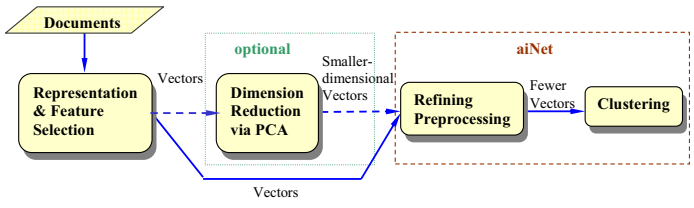


Figure 1: The framework of the proposed method.

The rationale of using the aiNet for document clustering is

SAC'05, March 13–17, 2005, Santa Fe, New Mexico, USA.

that it is capable of reducing data redundancy and obtaining a compressed representation of data [4]. In this way, more cohesive clusters are generated by the aiNet and thus resulting in better clustering results. The clustering results from the proposed method are compared with those from applying HAC and K-means directly. The comparison indicates that the proposed method, which inserts the refining preprocessing of the raw document collection before HAC or K-means, can reach higher clustering accuracy than those that apply HAC or K-means directly, especially with large numbers of documents. Since HAC and K-means are the two most commonly used methods for document clustering, only these two are used for comparison purposes. One can choose any other technique for the clustering part.

The rest of paper is organized as follows. Section 2 gives some basics about the immune system principles and then an overview of the aiNet. Section 3 discusses how the aiNet algorithm is further explored and applied to the problem of document clustering. The experimental results are explained in section 4. Section 5 includes the conclusion and some discussion of future work.

2. AINET - AN AIS METHOD FOR DATA ANALYSIS

The human body performs a variety of effective and powerful biological functions. Computer scientists have been exploring the mystery of these functions and applying their mechanisms to learning algorithms. Neural networks and genetic algorithms are two such families of algorithms. A third, yet a relatively new, family of biologically-inspired learning algorithms, dubbed artificial immune systems (AIS), began to draw people's attention. AIS are computational systems, inspired by theoretical immunology and observed immune system functions. They have been successfully applied to many application areas [2].

The aiNet is one such AIS approach to data clustering. For completeness, the immune system principles involved by aiNet are first summarized in section 2.1 and the aiNet algorithm is described in section 2.2.

2.1 Immune system principles

The immune system is a complex of cells, molecules and organs that aim to protect the body against infection. In the presence of infections, antigens, the substances capable of stimulating an immune response, are generated. The immune system usually produces a group of B-cells, which secrete antibodies. These antibodies can recognize and bind antigens and finally kill them. The affinity between an antigen and an antibody describes the strength of the antigen-antibody interaction, also referred as the affinity between the antigen and the B-cell. The larger the affinity between an antibody and an antigen, the tighter the antibody can bind the antigen. The body employs a group of immune mechanisms that can facilitate the B-cell generation to bind the antigens. The following subsections cover only a few principles that are exploited in the aiNet. A thorough description of immune principles can be found in [10].

2.1.1 Clonal selection and affinity maturation

The immune system randomly generates many B-cells. The B-cells with high affinity to antigens are cloned. These cloned cells can easily recognize and bind antigens, and are thus called *memory cells*. This cloning process of generating *memory cells* is called clonal selection. *Memory cells* have a longer life than normal B-cells and are thus useful when a similar infection occurs at a future time. The B-cells that have low affinity to antigens are either directly eliminated or mutated. The mutation process "edits" the antibodies bound on the surface of B-cells to obtain comparatively higher affinity to the antigens. This process of increasing affinity is called affinity maturation.

2.1.2 Immune network theory

The immune network theory indicates that the immune system involves not only the interaction of antibodies and antigens but also the interaction of antibodies with other antibodies. Cells can connect with each other to form a network representing an internal image of original antigens. The network can respond either positively or negatively. A positive response would result in cell proliferation, activation and antibody secretion. A negative response would lead to network suppression. Many different network models are presented in the literature but most of them can be generalized into the following formula [4]:

$$\text{RPV} = \begin{array}{c}\text{Influx of}\\ \text{new cells}\end{array} - \begin{array}{c}\text{Death of}\\ \text{unstimulated}\\ \text{cells}\end{array} + \begin{array}{c}\text{reproduction of}\\ \text{stimulated cells}\end{array},$$

where RPV stands for the rate of the population variant of the network.

2.2 The aiNet algorithm

In the aiNet [4, 3] algorithm, each data point is treated as an antigen. The algorithm evolves a population of antibodies based on the immune network theory, clonal selection and affinity maturation. These antibodies form a network, which can represent the antigens in a compressed way. The constructed network is called an aiNet model. Eventually clusters are automatically generated via HAC.

The procedure of evolving antibodies (Ab) to represent antigens (Ag) can be explained as follows. First randomly generate a set of Abs and put them into an empty memory matrix called M, and then follow the steps below:

(1) Affinity calculation: Calculate the affinity between the current Ag and each Ab from M. **(2)** Clonal selection: Select a subset of Abs with the highest affinity and clone them. The clone size is proportional to the affinities of Abs. That is, the higher the affinity is, the more Abs are cloned. **(3)** Affinity maturation: Mutate each Ab toward the current Ag with a rate inversely proportional to its affinity. As a result, the Abs with higher affinities experience comparatively smaller mutation. **(4)** Reselection: Calculate the affinity between each Ab and the current Ag; Reselect a subset of Abs with highest affinity and remove the Abs with low affinity to the current Ag. **(5)** Network suppression: Remove redundant Abs and insert the resulting Abs into M. **(6)** Repeat (1)-(5) for each Ag. The memory matrix M would eventually contain the *memory cells*, i.e., the Abs that bind the Ag closely for each Ag. **(7)** Suppress M: Remove redundant Abs in M to maintain an appropriate size. **(8)** Add a set of new randomly generated Abs into M. **(9)** Repeat (1)-(8) until a pre-defined number of iterations are reached.

After the antibodies are finally built, clusters are detected from these antibodies via HAC.

There are four tunable parameters for the aiNet. The analysis of these parameters for the document clustering experiments is covered in section 4.

n_s: the number of Abs selected for cloning in step **(2)**;

σ_s: the suppression threshold for step **(5)** and **(7)**, which defines the threshold to eliminate redundant Abs.

ζ: the percentage of reselected Abs for step **(4)**;

σ_d: the death rate, which defines the threshold to remove the low-affinity Abs after the reselection for step **(4)**.

In summary, the aiNet constructs a network of antibodies to represent the original antigens. The rate of antibody population variation of the network is proportional to the novel antibodies that are added in at each iteration (see step **(8)**) minus the death of low-affinity antibodies (the suppression in step **(5)** and **(7)**), plus the reproduction of high-affinity antibodies (the clone process in step **(2)**). This rate follows the RPV formula of the general immune network model. Together with clonal selection and affinity maturation, an aiNet model is eventually generated to represent the internal image of the original data. This internal image is capable of describing the distribution of clusters with less redundant and noisy information.

3. DOCUMENT CLUSTERING

To perform the task of document clustering, it is necessary to represent a document in a concise and identifiable format/model. Usually, each document is converted into an L-dimensional vector, where L is determined by the vocabulary of the entire document set. A collection of n documents then becomes an $n \times L$ matrix A, where L is the size of the lexicon. We use 0-1 vector representation in this paper, where the value of $A(i, j)$ is 1 if document i contains word j, otherwise 0.

3.1 Feature Selection

As mentioned above, the dimension L is the number of the total words appearing in the document set. L tends to be large even with a small set of documents. Instead of using all the words, l best words can be selected for clustering, where $l << L$. This can lead to significant savings of computer resources and processing time. It is called feature selection because each word is considered as a feature.

In the proposed clustering method, the following equation is used for feature selection. It evaluates the quality of a word w:

$$q(w) = \sum_{i=1}^{N} f_i^2 - \frac{1}{N}\left[\sum_{i=1}^{N} f_i\right]^2 . \qquad (1)$$

Here f_i is the frequency of word w in document d_i and N is the total number of documents. In [5], this method was used to choose 15% of words in the experimental document set while still keeping almost the same clustering results. The quantity $q(w)$ is proportional to the word frequency variance. By using this rating, words that are not uniformly distributed in the document set get higher quality.

3.2 Using the aiNet for Document clustering

After feature selection, each l-dimensional vector representing a document is created and treated as an antigen in the aiNet. By following the steps in section 2.2, aiNet generates a set of antibodies to represent the original antigens via an evolutionary process. These antibodies are the vectors containing the same features. The proposed method then detects clusters among the constructed antibodies via HAC or K-means. HAC is originally used in the aiNet algorithm (see section 2.2). K-means is added as another choice for clustering antibodies. The clusters of the antibodies are detected either via HAC or K-means, but the clusters of the original antigens remain unknown, i.e. it is still unknown as to which document belongs to which cluster. In order to obtain the cluster information of each document, the aiNet algorithm is modified to keep track of the antigens that are bound to each constructed antibody in the last iteration. In this way the cluster of a document is exactly the cluster of the antibody that the antigen is bound to. The entire procedure including antibody construction and clustering is referred to here as aiNet_HAC or aiNet_K-means, depending on the clustering method it uses. This document clustering procedure is summarized in the following chart. The details such as clonal selection and affinity maturation can be found in section 2.2.

aiNet_HAC or aiNet_K-means for Document Clustering

1. *M = [];*
2. *Convert n documents into n Ags via document representation and feature selection (each Ag is a l-dimensional 0-1 vector);*
3. *Randomly generate k Abs and put them into M;*
4. *Repeat for n_i iterations {*
  Repeat for each Ag {
    Calculate the affinity between the current Ag and each Ab from M;
    Clonal selection for M;
    Affinity maturation for M;
    Calculate the affinity between each Ab and the current Ag;
    Reselect a subset of Abs with highest affinity;
    Remove the Abs with low affinity to the current Ag;
    Suppress M;
  }
  IF the last iteration Then
    Suppress M while making the Ags of the removed Abs bind the existing Abs;
  Else
    Suppress M;
  Add k new randomly generated Abs into M;
}

5. *Cluster M which contains n' Abs via HAC or K-means;*
6. *Check the Ags of each Ab in M to obtain each Ag's cluster;*

The Principal Component Analysis (PCA) [7] is introduced to achieve a degree of dimensionality reduction before evolving the antibodies. The function of PCA is to reduce the dimension of the original vectors while preserving the variance-covariance structure of them. Usually, the resulting first few dimensions would account for a large proportion of the variability. For the purpose of this paper, if n documents are to be clustered, an $n \times l$ matrix is generated after the feature selection process. Here l is the number of the words with the highest quality in Eq.(1). Before directly set these n l-dimensional antigens as the input for the aiNet, PCA is used to reduce l into a much smaller number (say, 20) while still preserving about 65% of the information of the original document matrix (calculated by the percentage of the explained variability). Also, some noise information is removed to obtain better clustering results because the data not contained in the first few components may be mostly due to noise. Therefore, the $n \times l$ matrix is converted into an $n \times 20$ matrix via PCA. These n 20-dimensional vectors are taken as the input (antigens) of the aiNet algorithm. Via the aiNet a compressed representation, i.e., n' 20-dimensional antibodies, are generated to represent the original n antigens ($n' < n$) and then clustered. Thus the role of PCA is to compress the columns of the $n \times l$ matrix and the role of

aiNet is to compress the rows of the matrix. The remaining clustering process is the same as aiNet_HAC and aiNet_K-means. This procedure, which combines PCA, is referred to here as $aiNet_{pca}$_HAC or $aiNet_{pca}$_K-means depending on the clustering method it uses. The procedure with PCA inserts the PCA step "Reduce the dimension of the *Ags* via PCA: $l \rightarrow 20$;" between step 2 and step 3 of the procedure without PCA.

4. EXPERIMENTAL RESULTS

Experiments are conducted on the 20 Newsgroup data set [1]. This data set contains about 20,000 documents on different subjects from 20 UseNet discussion groups. Several subsets of documents with various degrees of difficulty are chosen. The details of the subsets are given in Table 1. For example, the subset_1 contains 150 randomly selected documents from each of the news groups sci.crypt and sci.space. Our preprocessing includes removing words appearing in a standard stop-list and skipping headers. The stop-list contains common words such as "the" and "of" that contribute nothing to the document subject.

Table 1: Some details on Datasets used.

Dataset	Topics included	#docs per group	Total #docs
subset_1	sci.crypt, sci.space	150, 150	300
subset_2	sci.crypt, sci.electronics	150, 150	300
subset_3	sci.space, rec.sports.basketball	150, 150	300
subset_4	talk.politics.mideast, talk.politics.misc	150, 150	300

By varying the four tunable parameters (n_s, σ_s, ζ, σ_d) mentioned in section 2.2, we obtained the best parameter setting for aiNet_HAC and aiNet_K-means – (2, 0.7, 0.1, 4), and the best parameter setting for $aiNet_{pca}$_HAC and $aiNet_{pca}$_K-means – (2, 0.07, 0.1, 4). The most essential parameter σ_s controls final network size and is responsible for the network plasticity. The different dimensionality of the antigens in these procedures with and without PCA results in the different values of σ_s. The parameters n_s and ζ can adjust the network size in a small degree based on the selected σ_s and make the final network as small as possible. σ_d is responsible for eliminating the antibodies with low antigenic affinity. It is only useful in the first iteration of the evolving process. These three less important parameters are the same with any of the aiNet procedures.

Table 2 displays the results for 6 clustering procedures: aiNet_HAC, $aiNet_{pca}$_HAC, HAC, aiNet_K-means, $aiNet_{pca}$_K-means and K-means. Two metrics are used to evaluate the clustering quality: accuracy (Acc.) and F-measure (F-mea.). Accuracy is defined as the percentage of correctly classified documents. The F-measure is another metric used in text mining literature for document clustering [9]. It combines the concepts of precision and recall.

Table 2 shows that the clustering results are significantly improved when using the aiNet (any of the aiNet procedure with or without PCA). The $aiNet_{pca}$_HAC ($aiNet_{pca}$_K-means) performs almost the same as aiNet_HAC (aiNet_K-means) and sometimes better, which shows that the aiNet with PCA can retrieve better or at least comparable clustering results while having less time complexity than that without PCA. The only case that the proposed method does not perform very well is with the subset_4. This is because subset_4 contains documents with relatively similar subjects. This causes the antibodies of a cluster to bind some borderline antigens from another cluster and thus results in the wrong classification of these antigens. This misbinding finally results in not-so-good clustering results.

The aiNet can obtain better clustering results than those without using it because it reduces data redundancy, thus generating more cohesive clusters. To prove this, i.e., to prove that the antibodies generated from the aiNet process form more cohesive clusters than the antigens, two criteria, namely homogeneity (H) and separation (S) are introduced:

$$H = \frac{1}{n}\sum_{i=1}^{n} dist(d_i, center_k)$$

$$S = \frac{1}{\sum_{i\neq j}|C_i|\cdot|C_j|}\sum_{i\neq j}|C_i|\cdot|C_j|\cdot dist(center_i, center_j)$$

They are two mathematical measures of the quality of the clusters. The metric H is calculated as the average distance between each data point and the center of the cluster it belongs to. The metric S is calculated as the weighted average distance between cluster centers. The metric H reflects the compactness of the clusters while S reflects the overall distance between clusters. Either a decrement in H or an increment in S brings the better quality of clusters. Figure 2 shows that the aiNet results in a smaller H (i.e. more compact clusters) with all the four document sets. There is only a small decrement in S. The impact of increasing in the compactness (i.e., H) exceeds the impact of a decrement in S, which makes the overall clustering result better.

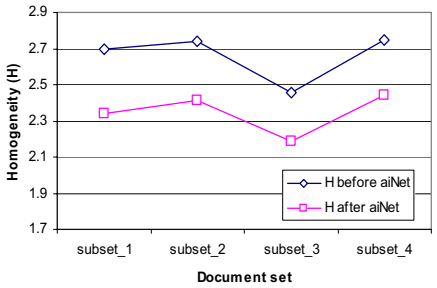
Figure 3 shows the clustering accuracies with different sizes of document sets. All the documents are selected from two news groups – *sci.crypt* and *sci.electronics* but with different number of documents. Subset_A randomly selects 80 documents from each of the two news groups making a total of 160 documents. Subset_B randomly selects 150 documents from each, thus making 300 documents. Similarly, subset_C randomly selects 300 documents from each, making 600 documents. The results indicate that the aiNet approach (with or without PCA) did improve clustering results when the size of the document set is large.

5. CONCLUSION AND DISCUSSION

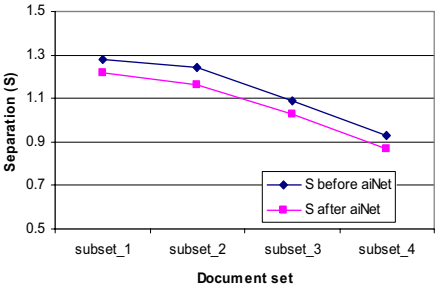
This paper presented a novel approach to document clustering based on the aiNet, an immune-system-based tool for data analysis. It first builds a compressed internal image of the original document collection and then automatically classifies them via classical clustering techniques. This approach was empirically tested with the 20 Newsgroup data sets. The experiments show that the refined internal image via the aiNet results in more compact clusters, and thus is capable of obtaining better clustering results than directly clustering the raw document set. PCA is also introduced into this approach to reduce the time complexity and to remove noise. The results show that by integrating PCA, our approach can cluster documents more accurately than without using PCA. Finally the experimental results also indicate that our approach is especially good for large-sized document sets that contain data redundancy and noise.

Table 2: Clustering results for different algorithms.

Method/Data	subset_1		subset_2		subset_3		subset_4	
	Acc.	F-mea.	Acc.	F-mea.	Acc.	F-mea.	Acc.	F-mea.
aiNet_HAC	0.817	0.810	0.687	0.640	**0.737**	**0.718**	0.59	**0.641**
$aiNet_{pca}$_HAC	**0.820**	**0.815**	**0.750**	**0.735**	0.730	0.715	0.60	0.640
HAC	0.500	0.665	0.557	0.654	0.723	0.700	**0.610**	0.631
aiNet_K-means	0.813	0.807	0.657	0.628	0.630	0.630	0.583	0.639
$aiNet_{pca}$_K-means	**0.840**	**0.836**	**0.693**	**0.661**	**0.660**	**0.631**	0.587	**0.646**
K-means	0.777	0.794	0.580	0.580	0.507	0.513	**0.597**	0.624



(a) Homogeneity



(b) Separation

Figure 2: Homogeneity and Separation for the data before and after the aiNet algorithm.

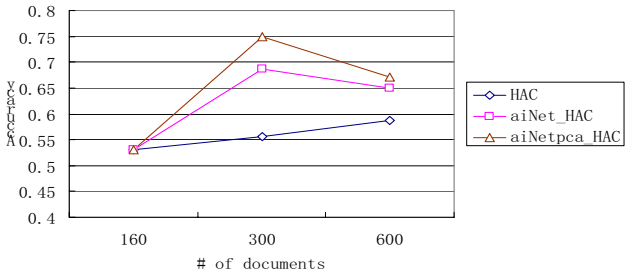


Figure 3: Clustering Accuracies with different size of document set.

Some interesting problems are left for future research: (1) Time complexity: Although PCA is introduced to reduce the processing time, the complexity of the aiNet is still quite large compared to other direct clustering techniques. Thus it is necessary to look for methods that reduce time complexity further so as to make it scalable with the size of the document collection. (2) Incremental documents clustering: The information environments tend to be dynamic and it is desirable to have an adaptive clustering method to deal with continuously growing document set. (3): Semi-supervised learning: It would be of some benefit if this approach can be modified into semi-supervised document clustering requiring only a small number of training documents, yet achieving better clustering accuracy.

6. ACKNOWLEDGMENTS

Work reported in this paper is supported in part by the AFOSR grant FA9559- -4-1-0159.

7. REFERENCES

[1] 20 newsgroup data set. http://people.csail.mit.edu/ jrennie/20newsgroups.

[2] L. N. de Castro and J. Timmis. *Artificial Immune Systems: A New Computational Approach.* Springer-Verlag, London. UK, 2002.

[3] L. N. de Castro and F. J. V. Zuben. An evolutionary immune network for data clustering. In *IEEE Brazilian Symposium on Artificial Neural Networks.*, pages 84–89, 2000.

[4] L. N. de Castro and F. J. V. Zuben. *AiNet: an Artificial Immune Network for Data Analysis.* Idea Group Publishing, 2001.

[5] I. Dhillon, J. Kogan, and C. Nicholas. Feature selection and document clustering. *Survey of Text Mining*, pages 73–100, 2004.

[6] K. Eguchi. Adaptive cluster-based browsing using incrementally expanded queries and its effects. In *ACM SIGIR 99*, 1999.

[7] I. T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, second edition, 2002.

[8] G. Muresan, D. Harper, and G. Mechkour. Webcluser, a tool for mediated information access. In *ACM SIGIR 99*, page 337, 1999.

[9] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, 2000.

[10] J. Timmis. *Artificial immune systems: A novel data analysis technique inspired by the immune network theory.* PhD thesis, 2000.

[11] F. Walls, S. S. H. Jin, and R. Schwartz. *Topic Detection in Broadcast News.* GTE/BBN Technologies, 1999.

[12] O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. In *ACM SIGIR 98*, pages 46–54, 1998.