Developing E-Mail Classification by using Artificial Immune System (AISEC)

Lwin May Thant, Daw Win Mar
Computer University (Maubin)
cupid.ucs@gmail.com

Abstract

Electronic mail is one of the primary applications of the Internet. With the increase in the use of email on the Internet, an email classification technology is increasing. In this paper, the immune based algorithm called Artificial Immune System for Email Classification (AISEC) that is capable of continuously classifying electronic mail as interesting (or) un- interesting without the need for re-training. The natural immune system exhibits many properties that are of interest to the area of web mining. Of particular interest is the dynamic nature of the immune system when compare with the dynamic nature of mining information from the web. By using AISEC algorithm, this paper is mining message that interesting and useful for every mail user depend on the using types of email. It can save a lot of time that the user need for reading and processing email.

Keywords: email, interesting, uninteresting, AISEC, mailbox.

1. Introduction

Web mining is moving the World Wide Web towards a more useful in which user can quickly and easily find the information the need. It includes the discovery and analysis of data, documents and multimedia from the World Wide Web. Web mining is an umbrella term used to describe three quite different types of data mining namely content, usage and structure mining(Chakrabarti,2003).Of these we are concerned with Web content mining, which Linoff and Perry (2001) define as "the process of extracting useful information from the text, images and other forms of content that make up the pages". The mining of textual data is a common web mining task; often for the purposes of information retrieval.

The content mining is becoming increasingly necessary as finding information on the Internet is almost impossible without automated assistance. The ultimate goal is to construct a system to mine from the web area that the user will find interesting. A system is to further work this area by the creation of an immune-inspired tool for mining interesting information from the web (Liu, Ma & Yu, *2001*).

At a high level, it is possible that we may take immune-inspired approach will have the ability to dynamically determine the interestingness of the retrieved result, where interestingness may consider them novel, surprising or unexpected [2].

In this paper we focus on the applying classification algorithm named "AISEC" (Artificial Immune System for Email Classification) is classified by depending on the subject and sender words and capable of continuous learning for the purposes of two-class classification of email. In the following paper is structured as follows: Section 2 is introduced motivation. Section 3 is explained basic concepts of AIS in detail. Section 4 is described the design and implementation of the system. Section5 is represented the conclusion and limitations.

2. Motivation

The number of people using email has grown dramatically over the last ten years and users receive increasing volume of email. An important one relates to the increasing amounts of time. In manual, when the mail arrives in the user mailbox, this user who reading the incoming mail [4]. If mail is the interesting (or) useful for the user, it is stored in the mailbox. Otherwise, it is deleted. So, responses for the interesting mails may delay and processing time consuming will cause. This paper intends to reduce this waste time and the mail user can get the interesting (or) useful mail massages in order to classify. Being based on an AIS algorithm, it is the most widely used in text mining and documents (or) email classification. They allow quick training, running (or) classification and can be easily extend to web-based information retrieval tasks. In AISEC algorithm, we do not pre-select a set of words from the training data; instead a selection is performed in a data-driven manner. Therefore, the mail user who interesting message in the short period can get.

3. Basic Concepts of AIS

Artificial Immune System (AIS) exhibits the desirable properties for a computational intelligence and memory. The natural immune system is based around a set of immune cells called lymphocytes and it is the manipulation of populations of these by various processes that gives the system its dynamic nature. Lymphocytes comprised of B and T-cells. From a data mining perspective, an important component of a natural immune system is a receptor.

These receptors are found on the surface of each lymphocytes and the binding of this receptor by chemical interactions to patterns. The main role of lymphocytes in AIS is encoding and storing a point in the solution space or shape space (Perelson & Oster, 1979). At the core of AIS is a set of immune cells, each described by a feature vector. Any object capable of binding to one of these receptors by chemical interactions is a called an antigen. Antigens are the objects to be classified and typically use the same representation as the immune cells that are a feature vector.

An affinity function may be defined to determine a measure of similarity between an immune cell and an antigen. If this affinity is higher than a threshold the antigen is said to be within the lymphocyte's recognition region of the immune cell as that the lymphocyte will recognize the antigen. The match between the receptor and antigen need not be exact and so when binding takes place it does so with a certain strength called an affinity. As lymphocytes may become activated by any antigen within this region a single lymphocyte may match a number of antigenic patterns, an important element of the noise tolerant nature of the immune system. When this binding takes place it stimulates an immune response from the lymphocyte and the cell begins to clone and mutate.

The cloning takes place with a rate proportional to affinity and mutation with a rate inversely proportional to affinity in a process called colonel section. Colonel section constitutes the core of the immune system's adaptation mechanisms. An adaptation process such as this is a common paradigm found in many evolutionary algorithms, but asexual reproduction and mutation with rate dependent on some fitness measure are an important difference between AIS and these others. A T-cell needs two signals to become activated. The recognition of an antigen by a T-cell is said to be signal one, the second signal or co-stimulation signal is a confirmation as is given by an antigen presenting cell. It has bound to a cell not properly presented, assumed to be part of the host and the T-cell is removed. [2]

3.1 AIS for E-Mail Classification

E-mail classification was important to produce a text mining system based on an immune inspired algorithm. This must be tested in a dynamic domain. The performance of our text mining system on the task of e-mail is to distinguish between e-mail the user would not be interested in and legitimate e-mail which to the user is important or interesting. The text contained in the e-mail is used for the purposes of classification and email is part of the Internet environment [2]. A further reason for this choice of testing scenario was that the problem of receiving uninteresting or junk e-mail is one faced by most that use e-mail on a day-to-day basis.

An AISEC algorithm for mining interesting information from the Web, and is performing a task similar to a spam filter. It has no special measures to cope with this asymmetric loss function. The penalty for misclassifying an interesting document when the final system is run is not nearly as severe as misclassifying an e-mail. The main difference is that we address a continuous learning scenario where the concept of what the user finds interesting will change over time and so may the content of uninteresting e-mails. This contrasts with the vast majority which are trained once and then left to run. AIS are that our e-mail classifier requires no specific feature selection mechanisms. In contrast, we do not pre-select a set of words form the training data; instead a selection is performed in a data-driven manner implicitly by the evolutionary operators [2].

3.2 AISEC Algorithm

Artificial Immune System for E-mail Classification (AISEC) seeks to classify an unknown e-mail into one of two classes. It does this by manipulating the populations of two set of immune cells. Each immune cell combines some features and behaviors from both natural B-cells and T-cells but for simplicity we refer to these as B-cells throughout. These two set consists of a set of naïve (sometimes called free) B-cells and a set of memory B-cells. Once the system has been trained, each B-cell represents an example of an uninteresting e-mail by containing words from that e-mail's subject and sender fields in its feature vector.

New e-mail to be classified by the system are considered to be antigens and so to classify an e-mail (antigen), it is first processed into the same kind of feature vector as a B-cell and then presented to all B-cells in the system. If the affinity between the antigen and any B-cell is higher than a given threshold, it is classified as uninteresting; otherwise it is allowed to pass to the user's normal inbox. If the antigen (e-mail) is classified as uninteresting, it will be removed to a temporary store. If the user deletes an e-mail from the temporary store, it is confirmed to represent an uninteresting e-mail. The B-cell that classified it as uninteresting is useful and is selected for reproduction. The constant reproduction combined with appropriate cell death mechanisms gives the AISEC algorithm its dynamic nature.

z

During design a number of special considerations were given to the specialist nature of the text mining domain. The incorporation of these considerations in the final algorithm served to further distance our system from other AIS. These design decision are discussed below.

Representation of one data class: In a Web-mining context, the number of documents a user finds interesting may be tiny compared with those a user finds uninteresting. B-cells therefore represent only the uninteresting e-mail class. This is a helpful simplification for the purposes of efficiency and more akin to the way the natural system works. Natural lymphocytes only encode possible antigenic patterns and everything else is assumed harmless.

Gene libraries: Two libraries of words, one for subject words and one for sender words are used. These contain words known to have previously been used in uninteresting e-mail. When a mutation is performing, a word from this library replaces a word from a cell's feature vector. Mutating a word in any other way, by replacing characters for example, would result in a meaningless string in almost all cases.

Co-stimulation: Uninteresting e-mail is not deleted but stored away. A B-cell must have become stimulated to classify this e-mail, so it can be assumed the first signal has already occurred. User feedback is then used to provide or not provide a signal. At a time of the user's convenience this store may be emptied. It will be these user actions that will drive a number of dynamic processes. If an e-mail is deleted from this store by the user, the system has performed a correct classification; the user really was not interested in that e-mail and so a co-stimulation signal has occurred. The cell is rewarded by being allowed to reproduce. If, on the other hand, the user does not delete the e-mail the algorithm has performed a misclassification, signal two does not occur and B-cells are removed appropriately.

Tow recognition regions: Around each B-cell is a recognition region within which the affinity between this cell and an antigen is above a threshold. It is within this region an antigen may stimulate another. A single region was found to be inefficient for both the triggering of evolutionary processes and classification. A smaller region, a classification region, was introduced for use in classification only.

Cell death processes: To both counteract the increase in population size brought about by reproduction and keep the system dynamic, cell death processed must be implemented. A naive B-cell has not proved its worth and is simply given a finite lifespan when created, although it may lengthen its life by continually recognizing new pieces of data confirmed as uninteresting. Memory B-cell may also die, but these cells have proved there worth and it can be hard for the system to generate clones capable of performing well. For this reason, unlike B-cell, memory cells are purged in a data driven manner. When a new memory cell, mc, is added to the memory cell set all memory cells recognizing mc have a stimulation counter reduced. When this count reaches zero they are purged from the system [3].

3.3 Notation in AISEC Algorithm

Before we begin, let us establish the following notational conventions:

Let BC refer to an initially empty set of naive B-cells

Let MC refer to an initially empty set of memory B-cells

Let Kt refer to the initial number of memory cells generated during training

Let Kl refer to the clone constant that controls the rate of cloning

Let Km refer to the mutation constant that controls the rate of mutation

Let Kc refer to the classification threshold

Let Ka refer to the affinity threshold

Let Ksb refer to the initial stimulation count for naive B-cells

Let Ksm refer to the initial stimulation count for memory B-cells

3.3.1 Representation

A B-cell receptor holds information that may be extracted from a single e-mail; this is represented as a vector of two parts (Figure1). One part holds words present in the subject field of a single e-mail and the second holds words present in the sender (return address) fields of that particular e-mail. The actual words are stored in the feature vector because once set, the vector will not require updating throughout the life of the cell. This can be contrasted to the common practice of using a vector containing binary values as the receptor, each position in which represents the presence or absence of a word known to the system. As words are continually being added and removed from the system, each cell's vector would have to be updated as appropriate when this action occurs. The two sub-vectors are unordered and of variable length. Each B-cell will contain a counter used for aging the cell that is initialized to a constant value on generation and decremented as appropriate. This counter may be re-initialized if the B-cell is added to BC [3].

B-cell vector = (subject, sender)
Subject = (word 1, word 2, word 3… word n)
Sender = (word 1, word 2, word 3… word n)

Figure 1: B-Cell Structures

3.3.2 Affinity Measure

z

The affinity between two cells measure the proportion of one cell's feature vector also present in the other cell. It is used throughout the algorithm and is guaranteed to return a value between 0 and 1. The matching between words in a feature vector is a case insensitive but otherwise requires an exact character.-wise match [5]. Given bc1 and bc2 are the cells we wish to determine the affinity between, the procedure may be outlined as follows:

```
PROCEDURE affinity (bc1, bc2)
IF (bcl has a shorter feature vector than bc2)
        bshort ←bcl, blong ←bc2
ELSE
        bshort ←bc2, blong ←bcl
count ←the number of words in bshort present
        in  blong
bs_len ←the length of bshort's feature vector
RETURN count/bs_len
```

Figure 2: Affinity Process

3.3.3 Algorithms and Processes

The AISEC algorithm works over two distinct stages: a training phase followed by a running phase. An overview of this algorithm is described in Figure 3.

```
PROGRAM AISEC
train (training set)
WAIT until (an e-mail arrives or a user action is intercepted)
  ag ←convert e-mail into antigen
      IF (ag requires classification)
         classify (ag)
            IF(ag is classified as uninteresting)
             move ag into user accessible
             storage
           IF(do you insert temporary
               e-mail to inbox)
              added to the user inbox
           ELSE  deleted ag
        ELSE allow e-mail to pass through
      ELSE move to the user inbox
```

Figure 3: AISEC overview

During the training stage the goal is to populate the gene libraries, produce an initial set of memory cells from training examples and produce some naive B-cells in the AISEC algorithm represent one class, only the entire training set, here called **TE,** contains only e-mails the user has positively selected to be uninteresting. When the training process, the procedure in Figure 4 may be used as follows:

```
PROCEDURE train (TE)
   FOREACH (te ∈TE)
         process e-mail into a B-cell
         add subject words and sender words to
            appropriate library
         remove Kt random elements from TE
            and insert into MC
        FOREACH (mc ∈ MC)
                 set mc's stimulation count to
                   Ksm
         FOREACH (te ∈ TE)
         set mc's stimulation count to
                  Ksb
           FOREACH (mc ∈ MC)
                 IF (affinity (mc, te) > Ka)
                   clones   ← clone- mutate
                             (mc,te)
```

Figure 4: Training Process

Now that the system has been trained it is available to begin the classification of unknown e-mail. During the running phase the system will wait for a new mail to be classified. To classify a new e-mail, an antigen, ag, is created in the same form as a B-cell, taking its feature vector elements from the information in the e-mail, then assigned a class based on the classification result. When classify a new email, the following procedure in Figure5 is used.

```
PROCEDURE   classify (ag) returns a classification
        for ag
FOREACH (bc ∈(BC ∪  MC))
              IF (affinity (ag, bc)) > Kc)
                   classify ag as uninteresting
                   RETURN
         classify ag as interesting
         RETURN
```

Figure 5: Procedure for Classification

The process of cloning and mutation which has been used throughout this section is detailed in Figure 6.bc1 is the B-cell to be cloned based on its affinity with bc2.kl and km are constants used to control the rate of cloning and mutation. The greatest integer smaller than or equal to the real-valued number x and num-clones, num-mutates must be integers [2].

```
PROCEDURE  clone-mutate (bcl, bc2) returns
               set of B-cells
       aff← affinity (bcl, bc2)
       clones ← ϕ
      num_clones ← ⌊ aff * Kl ⌋
       num_mutate ← ⌊ (1-aff) * bc's feature
                            vector length * Km ⌋
       DO (num_clones) TIMES
              bcx ← a copy of bcl
            DO (num_mutate) TIMES
                   p ← a random point in
                        bcx's feature vector
                   w ← a random word from
                   the appropriate gene library
                   replace word in bcx's feature vector
                   at point p with w
       bcx's stimulation level ← Ksb
             clones ← clones ∪ {bcx}
    RETURN clones
```

z

Figure 6: Cloning and Mutation Processes

4. Implementation of System design

The system can accept the email from the mail user of using this system. This system can use for every usage types of people that using email. This system is training phase followed by a running phase.

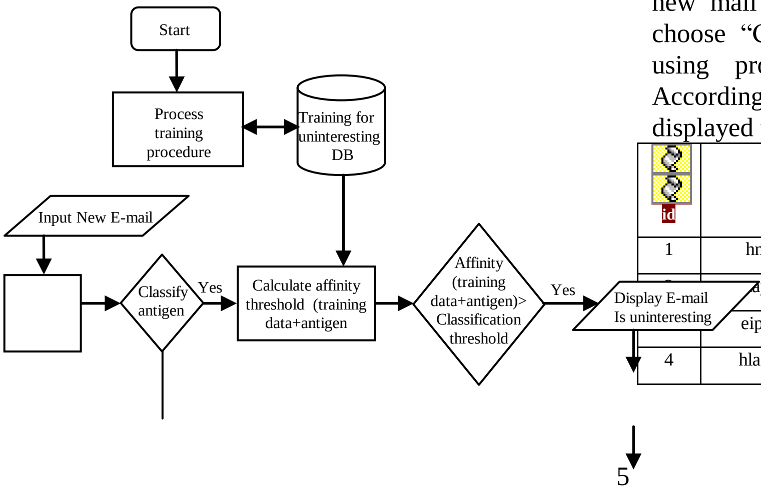
In the training phase, process the training procedure based on the positively selected by the user of uninteresting email and then training data is stored in the training database. Now that the system has been trained, this system wait for new mail arrive.

If new e-mail is arrived into the system and then new email is converting to antigen such as B-cell structure. And then, the mail user is asked that this mail is classify or not by the system. If new mail is not classified, this new mail (an antigen, ag) is directly added to the mailbox.

If the mail user is to classify an email (ag), then we compute affinity measure between training email (uninteresting) and new email (ag). If an affinity measure is less than a given classification threshold, display new email (ag) is interesting and insert into mailbox and the system is finished.

If an affinity measure is higher than a classification threshold, display new email (ag) is uninteresting and then removed to a temporary store. Then, this system is asked the user that this email (ag) is uninteresting but not useful or useful. If an email (ag) is useful for mail user, then this email (ag) is added to the user inbox and the system is finished.

If email (ag) is not useful, then email (ag) is deleted from a temporary store and the system is finished. In this way, the next incoming email is classified by the system but not against training process. The design of the system is described in the figure7.



Start
Process training procedure
Training for uninteresting DB
Input New E-mail
Classify antigen
Yes
Calculate affinity threshold (training data+antigen
Affinity (training data+antigen)> Classification threshold
Yes
Display E-mail Is uninteresting

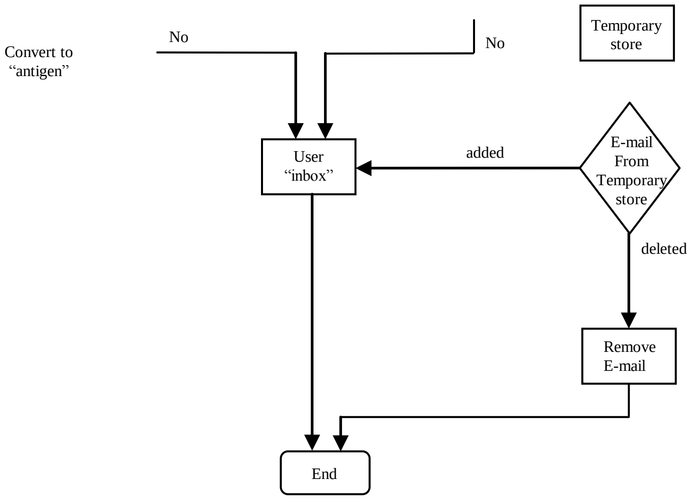


Figure 7: Detail Design of the System

4.1 Implementation of the System

This system is an implementation of E-mail classification with AISEC algorithm and used every email user. This system is implemented by C# programming language with mail server database. We were able to test the algorithm on a standard email 25 data set is used. Only the worlds contained in the subject and sender fields of the email were used, but the sender information also included the return address. Throughout the system, the above section process is used. The following figure is the process of our implementation.

Training Stage: When we start this system need to train the data set. Training data set consists of uninteresting mail that are trained depending on the positively selected by the using types of mail user. By clicking "Training "button in the following page, produce an initial set of memory cells and some naive B-cell. The training result is stored in the "Finish Training Table". This stage is used the training process in the above section of Figure 4. The following figure 8 is an email 25 data set for training this system.

Running Stage: After training stage, this system wait for new mail arrive. By clicking the "Show Current Email" button, show the address and subject word of sending new email. And then, the users choose the new mail is inserting into inbox or classify. If user choose "Classify" button, the system is calculated by using procedure for classification in Figure5. According to the classification result, the system is displayed the new mail is interesting or uninteresting.

id	address	subject
1	hninwai@gmail.com	Your Friend Photo
2	aythu87@gmail.com	I miss you
[illegible]	eiphyu87@gmail.com	Chat with photo
4	hlaingmin@gmail.com	Uninterruptible Power Supply

z

5	orlando88@gmail.com	Microsoft PowerPoint Book
6	kokothein@gmail.com	business manegement lecture notes book
7	ayekyaw@gmail.com	dental disease analysis paper
8	syithuwin@gmail.com	sea and storm subject paper
9	aungthwin@gmail.com	computer application and business management paper
10	doctormya@gmail.com	nature of wild animals book
11	thantun@gmail.com	myanmar history paper
12	aungkhinsint@gmail.com	healthy family book
13	nyankyaw@gmail.com	poison and food book
14	amayein@gmail.com	book shopping
15	goldenhousing@gmail.com	advertising for home sales book
16	phemyint@gmail.com	world boss men book
17	khinmaungtun@gmail.com	myanmar football history book
18	uhlabaw@gmail.com	geological analysis paper
19	doctormatinwin@gmail.com	bayintnaung naungyo fighting book
20	khinthawda@gmail.com	computer science changes paper
21	monmon@gmail.com	robot workers book
22	kyawthukhant@gmail.com	book of lady fashing
23	zawlwinoo@gmail.com	paper for current myanmar football
24	oasisstudio@gmail.com	mp3 and mp4 analysis paper
25	ucsy@gmail.com	misc paper analysis

Figure 8: Sample training data for uninteresting e-mail

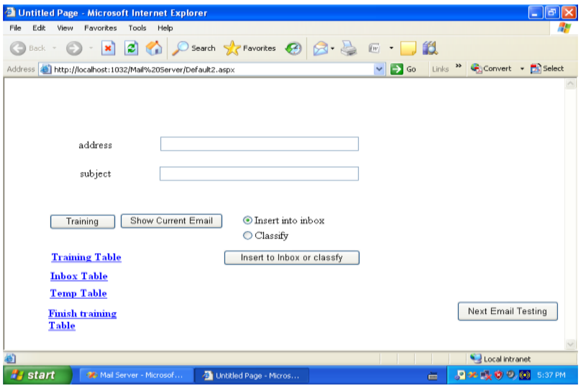

Untitled Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:1032/Mail%20Server/Default2.aspx
Go Links Convert Select
address
subject
Training
Show Current Email
Insert into inbox
Classify
Training Table
Insert to Inbox or classfy
Inbox Table
Temp Table
Finish training Table
Next Email Testing
Local intranet
start
Mail Server - Microsof...
Untitled Page - Micros...
5:37 PM

Figure 9: Loading Process for E-Mail Classification

5. Conclusion and Limitation

An immune inspired system is capable of the specialized task of document classification using a text-mining approach. An artificial immune system for web mining we have described a novel immune inspired algorithm is used classification techniques of email. This thesis develops a system that utilized AISEC algorithm and classify the mail messages. This email classification system will support to classify email of every users mail messages.

According to the user interesting, mail user can get the interesting (or) useful email without time-consuming with high accuracy. An increase in accuracy may be achieved by a change in the data stored in the B-cells feature vector such as measure of the relative importance of words and coupled with the necessary change in affinity function. In this system, unlike traditional single shot learning where there is need a fixed test set. Each time a new email is classified the algorithm can use the result of this classification to update it internal representation.

To purge the algorithm of cells which may match interesting email, the AISEC algorithm uses the two signals. Signal one is the antigen generated from the classified email has already stimulated a B-cell to have been classified. This signal one is used in this system. But signal two that comes from the user in the form of interpreting the user's reaction to this email is not used in this system. An AISEC algorithm is developing a web mining system to return mail information based on a measure of the user interestingness. So, this system can provide the quality of email classification.

References

[1] A. D. Pereslon and G. F. Oster
"Theoretic studies of colonel section: minimal antibody repertoire size and reliability of self- non-self discrimination"
Journal of Theoretical Biology, 1979

[2]A. Scime
"Web Mining: Applications and Techniques"
Idea Group Publishing, 2004

[3] A. Secker, A.A.Freitas, J.Timmis
"AISEC: an Artificial Immune System for E-mail Classification"
Paper presented at the First International Conference on Artificial Immune System (ICARIS), Canterbury, UK, 2002

[4] E. Crwaford, J. Kay and E. McCreath
"Automatic induction of rules for email classification"
Paper presented at the Australian Document Computing,2001

[5] J. Twycross and S. Cayzer
"An immune-based approach to document classification"
Information Infrastructure Laboratory, 2002