Incremental Immune-Inspired Clustering Approach to Behavior-Based Anti-Spam Technology

Xiao-bing LIU, Nan ZHANG

CIMS Cent., DaLian University of Technology , DaLian , 116024,China
E-MAIL: nand@sohu.com

Abstract

Facing new type of challenge which maintain clusters in a dynamic web environment with a high volume of updates and costly re-clustering, the paper describes a novel behavior-based anti-Spam technology based on incremental immune-inspired clustering algorithm. we use an "internal image" network to represent the input data set in order to reduce data redundancy, whilst at the same time extracting relevant information from the data set, It is difference with traditional clustering methods and it's ability is to deliver the most relevant Spam from the collection of all Spam that is reported by members without having to perform complete re-clustering, Experimental evaluation shows that the novel approach provide significantly faster data summarization than completely re-clustering and the technology is reliable, efficient and scalable. Because no single technology can achieve one hundred percent Spam detection with zero false positives, it should be used in conjunction with other filtering systems to minimize errors.

Keywords: Incremental learning, clustering, Artificial Immune Network, anti-Spam,

1.Introduction

Spam is becoming more inconvenient, annoying and wasteful of computer resources, and personalized anti-Spam filters of email client application based on Content filters have now become the standard for Spam filter. A variety of algorithms and rules have been implemented by rule-based filters, nearest- neighbor classifiers, decision trees and Bayesian classifiers [1]. But, the widespread adoption of personalized text classification has prompted the Spam senders to develop a technique that involves adding words and phrases to the body of an e-mail [2]. The intent of this attack is to cause the user's classier to falsely accept single Spam e-mail, which can degrade classier performance.

E-mails are filtered inconsistently across different users according to whether they are of interest to the user. Since users are not uniform in their desires or business needs, a good anti-Spam filtering system should take the different users' needs and desires into consideration and influence the same decisions and the behavior. The thesis underlying our research is that email server can dynamically adapt and collaboratively deliver the same Spam behavior-based patterns based on the feedback from behavior-based patterns among individual client users in the same e-mail server. Server

Spam filter, like a firewall, can protect the network from fighting Spam by blacklisting and related technology.

But, these collection data of user activities always arrives as multiple, continuous, rapid and time varying flow. Applications of clustering in such web data environment are now facing a new type of challenge: maintaining clusters in a dynamic environment with a high volume of updates without frequently performing complete and costly re-clustering.

Incremental methods are of great interest to cope with this challenge. Incremental algorithm for clustering in particular is difference with traditional clustering methods [3] and it allows dynamic tracking of the ever-increasing large scale information without having to perform complete re-clustering. There are several incremental clustering algorithms that do not use the data summarization technique but attempt to directly restructure the clusters to reflect the dynamic changes of the dataset, Chen et al. [4] propose the incremental hierarchical clustering algorithm GRIN which is based on gravity theory in physics. Ester et al. [5] present a new incremental clustering algorithm called Incremental DBSCAN which based on the DBSCAN algorithm [6]. Widyantoro et al. [7] present the agglomerative incremental hierarchical clustering (IHC) algorithm that also utilizes a restructuring process while preserving homogeneity of the clusters. Charikar et al. [8] introduce new deterministic and randomized incremental clustering algorithms while trying to minimize the maximum diameters of the clusters . M. Charikar[9] is the first to examine clustering Web content for efficient replication, and use both replication performance and stability as the metrics for evaluation of content clustering .

In this paper we present an incremental clustering algorithm based on artificial immune network which is capable of continuously identifying similar groups of SPAM. Artificial Immune System is a new, biologically inspired, paradigm of information processing. The immune network model discussed in this paper was introduced by de Castro and Von Zuben[10], and named aiNet (Artificial Immune NETwork). The main role of the standard adaptive algorithm proposed for the aiNet was to reduce data redundancy, whilst at the same time extracting relevant information from the data set, such as the spatial distribution of the inherent data clusters. The network cells within aiNet are represented in a space of same dimension as the input data, i.e. no dimensionality reduction is performed, but the network size is controlled based upon the immune network dynamic and metadynamic processes [11]. The network cells represent an “internal image” of the input data set, Consider a dataset to represent an antigen universe: a single item of data in the data set represents an antigen that must be recognized. An antibody produced by the artificial immune network recognizes a set of antigens in the antigen universe, Assuming all antigens in the universe can be recognized by the antibody set, then the number of antibodies present determines the generality/specificity of the clusters; a small number of antibodies will result in few clusters, and therefore each cluster represents a very general description of the data. As the number of antibodies is increased, the specificity of the cluster and hence the concept it represents also increases. New data arriving in the data-base is continuously presented to the system, which triggers the antibody set to adapt to the new dataset, either by adapting existing antibodies, or creating new ones. This is exactly analogous to the primary response in the biological immune system.

The rest of the paper is organized as follows: In section 2, we describe the Classical artificial immune network , and then we focus our discussion on the Incremental algorithm for clustering , In section 3, we describe the behavior-based characteristics of Spam and our discussion on behavior-based algorithm. In section 4, we present the results of a performance study that investigates the effectiveness of proposed techniques. We conclude the paper in section 5.

2: Incremental Immune-Inspired Clustering approach Length

2.1. Immune-Inspired Clustering Algorithm

Artificial Immune System is a new, biologically inspired paradigm of information processing. As a parallel and distributed adaptive system, it exhibits the following points of strength: recognition, feature extraction, diversity, learning, memory, distributed detection, and self-regulation; Three immunological principles are primarily used in a piecemeal in AIS methods, including the immune network theory, the mechanisms of negative selection and the clonal selection principles. The technique we use in our system is based on the aiNet (Artificial Immune NETwork) [12,13]. It is an artificial immune system inspired by the immune network theory, originally proposed by [14]. It is an iterative clustering algorithm that performs data compression through a pattern recognition process. aiNet is an artificial immune network model originally developed to perform automatic data compression. Combined with graph theoretical and statistical clustering techniques, aiNet is a powerful data clustering and classifying tool.

Classical ainet algorithm :

- Initialization: create an initial random population of network antibodies;
- Repeat for each antigenic pattern, do:
  1.Clonal selection and expansion: for each network element, determine its affinity with the antigen presented. Select a number of high affinity elements and reproduce (clone) them proportionally to their affinity;
  2.Affinity maturation: mutate each clone inversely proportional to affinity. Reselect a number of highest affinity clones and place them into a clonal memory set;
  3 Clonal interactions: determine the network inter-actions (affinity) among all the elements of the clonal memory set;
  4 Clonal suppression: eliminate those memory clones whose affinity is less than a pre-specified threshold;
  5 Metadynamics: eliminate all memory clones whose affinity with the antigen is less than a pre-defined threshold;
  6 Network construction: incorporate the remaining clones of the clonal memory with all network antibodies, resulting in a matrix M of memory antibodies;
  - Network interactions: determine the distance between each pair of network antibodies and store these data in a matrix D;
  - Network suppression: eliminate all network antibodies whose affinity is less than a pre-specified threshold;
  - Diversity: introduce a number of randomly generated cells to the network;
- Repeat until a pre-specified number of iterations is performed .

The key role of the algorithm is by reproducing (cloning) those cells capable of appropriately recognizing specific pathogens. During the proliferative phase of the immune cells, they are subjected to a controlled mutation event with high rates, termed somatic hypermutation. Those mutated offspring cells that have increased their capability of recognizing a specific pathogen are then selected for survival and further reproduction. This whole mutational process followed by selective events is called affinity maturation of the immune response, because it allows the immune system to increase its capability to recognize (affinity with) pathogens. A population of immune cells that reproduce under the effects of mutation and then suffer (natural) selection is a remarkable example of the evolutionary nature of an adaptive immune response. There are two types of stopping iteration; one is maximum number of generations; another is Stop the iterative process .

Now ,The M represent an "internal image" of the input data set, then we use the antibodies number of M to determine the generality/specificity of the clusters; a small number of antibodies will result in few clusters, and therefore each cluster represents a very general description of the data. therefore it became necessary to use additional tools in order to automatically identify and separate clusters in this network of cells. minimal spanning tree (MST) shall be a useful mechanism with which to automatically detect and separate the network clusters[15]. the MST is very effective at processing networks produced from aiNet, as aiNet positions network cells in appropriate locations within the space. The MST is a graph-theoretic technique which determines the dominant skeletal pattern of a point set by mapping the shortest path of linear, nearest-neighbour connections. the MST network represents a cumulative statistical summary of the spatial characteristics which underly such graphs and provide a visual, geometric summary of fungal sporogenesis.

2.2. Incremental Clustering Algorithm

The original aiNet model suffers from the lack in a dynamic web environment with a high volume of updates and costly re-clustering, In this section, we examine how to incrementally add new data to existing clusters, the novel algorithm aims at enhancing the incremental clustering capability of aiNet, such that clusters can be detected within a previously defined cluster. we take the centroids of each cluster to create clusters incrementally and dynamically based on the similarity between new data and the centroids of each cluster.

Algorithm(incremental clustering algorithm for aiNet,):

Input :

ΔX is increment data set of X

M: is memory cells matrix last time

S is matrix containing the centroids of each cluster last time

d min-threshold

Output:

$M' = M + \Delta M$ ΔM is increment data set of M

k' is new number of clusters of $X + \Delta X$

S' is new matrix containing the centroids of each cluster of $X + \Delta X$

`Procedure dyainet(`ΔX`,M,K,S)`

Step1: define M as globel $Ab \leftarrow M$ // the set of memory cells M last time will be used as network antibodies this time .

Step2: For each $x_i \in \Delta X$: // Repeat for each antigenic pattern, do:

Step3: $D = \{ d(x_i, y_j) \}$ *where* $x_i \in \Delta X \quad y_j \in S$

Step4: IF $D > d\min - threshold$ then run step4,step5,step6 of algorithm1

Step5: else $M \leftarrow M + x_i$

Step6: Repeat

Step7: use of the minimal spanning tree (MST) to calculate k' and S' again.

There are two main strategies to address the problem of incremental clustering in a data set environment. When new data are added to the data collection, ΔX is increment data set of X , and $|S| = k$ is matrix containing the centroids of each cluster. Let $D = \{ d(x_i, y_j) \}$ where $x_i \in \Delta X \quad y_j \in S$ be the similarity distance between new data and the centroids of each cluster.

If $D >= \min - threshold$ then re-extracting relevant information from $X' = X + \Delta X$ //

else

$M \leftarrow M + x_i$ // insert new data into memory cells matrix

Endif

3. Behavior-Based Characteristics of Spam

Compared with popular personalized anti-Spam filtering system based on text classification technology, the goal of our behavior-based anti-Spam technology at email services also give Spam a "score" that can be used as input to cluster. We shall analyze the message's headers, the body of the message and its structure, then give our numeric score that indicates how likely they are.

1.Sender IP address and SMTP id number :

The received lines in email headers contain the list of IP addresses that email has flowed through. As it is passed from one server to another, many of these lines can be faked. The key line is the first one internal to the recipient's organization, which gives the IP address of the outside machine delivering the message across the internet to the recipient's organization. This line can be trusted. We will refer to this line as the "first internal line." Identifying this line in clients is the goal of this paper. We can keep looking through the list until we reach a machine listed in the MX record. If the machine listed in the MX record received from an external sender, we can be very confident that we have found the sender.

IP address:

$$X = \{x_4, x_3, x_2, x_1\} \quad x_1, x_2, x_3, x_4 \in Integer$$

$$Score\ \Psi = \sum_{i=1}^{4} x_i \times 10^{i-1}. \tag{1}$$

SMTP id number:

$$Y = y_n y_{n-1} \cdots y_k \cdots y_3 y_2 y_1 \quad y_n y_{n-1} \cdots y_k \cdots y_3 y_2 y_1 \in Text$$

$$Score\ \Psi = size(Y) \times 10^6 + \sum_{k=1}^{8} ASCIIvalue(y_k) \times 2^{k-1}. \tag{2}$$

2. URL link or Reply email Address in Spam messages

Spam offenders use special e-mail addresses or web pages for "mailto" links. Messages inform the readers to reply to the mail with a subject of "REMOVE", and there is the major "call-to-action" for Spammers.

$$N = \cdots @ \alpha.\beta \cdots \qquad N = http://www.\alpha.\beta \cdot \cdot \qquad \alpha = z_n z_{n-1} \cdots z_k \cdots z_3 z_2 z_1$$

$$\beta = z_n z_{n-1} \cdots z_k \cdots z_3 z_2 z_1$$

$$\omega 1 = size(\alpha) \times 10^3 + \sum_{k=1}^{8} ASCIIvalue(z_k) \times 2^{k-1} \tag{3}$$

$$\omega 2 = size(\beta) \times 10^3 + \sum_{k=1}^{8} ASCIIvalue(z_k) \times 2^{k-1} \tag{4}$$

$$Score \qquad \Psi = \omega 1 \times 10^2 + \omega 2 \tag{5}$$

A Spam that users feed back to e-mail server is a text file. We should extract the characteristics of Spam based on equation (1,2,5).

4. Experiments and Results

We have developed a tool to automatically catch and calculate the "score" of Spam based on section 3, hen we shall use them as input to cluster same Spam. In this section, we mainly present the results of a performance study that investigates the efficiency of our Incremental algorithm for clustering. The results show that our new method for incremental clustering is suitable to be used in web dynamically changing environment, All the experiments were done on a PC with one 1.6GHz Intel Pentium 4 CPU, 256M host memory, running windows XP ,MATLAB Release 13

We create databases using synthetic data to simulate web dynamic data environment where there are random insertions to some clusters in the database, while other various clusters appear, disappear, and move. We populate our databases with 100 to 5,000 points to simulate a reasonable average of the data set size due to the limitation of our computer .

The real data stream for our experiment is the April 2003 standard testing corpus available by the SpamAssassin web site [16]. The set of 1400 messages was shuffled

fourteen times, with each shuffle providing an effectively random sequencing of the messages.

After some preliminary tests with the algorithm, the aiNet has some parameters to be defined in order to run the algorithm, though no exhaustive search of parameters or sensitivity analysis is performed. Table 1 shows the combination of parameters used.

Table 1. Values used for the aiNet parameters

	Paramete	Value	Meaning
1	n	4	best-matching cells taken for each Ag
2	N	10	clone number multiplier
3	qi	0.2	percentile amount of clones to be Re-selected
4	gen	40	maximum number of generations;
5	tp	1	pruning threshold;
6	ts	0.1	Suppression threshold
7	mi	4	learning (hypermutation) rate
8	sc	0.01	pre-specified number of iterations

In the first series of simulations, we first compare the Effectiveness of Proposed incremental Algorithm for clustering with re-clustering, We perform experiment where we demonstrate that if we use a simple data set that consists of five clusters As Figure 1 shows , and used re-clustering algorithm to identify the centers of the clusters. The results for some of these experiments are contained in Figure 1.

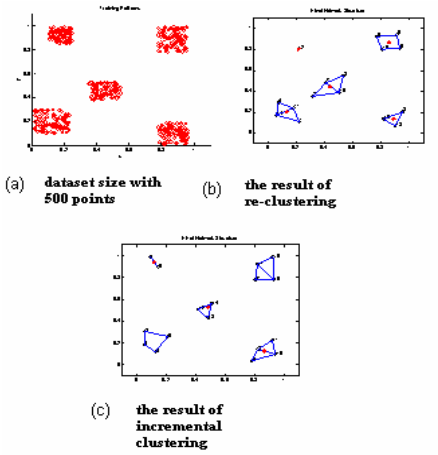
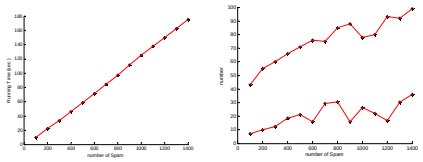


Figure **1.** the result of re-clustering the incremental clustering algorithm

Then dataset were divided into 5 subset , and used our incremental algorithm for clustering to identify the centers . results indicate that the incremental approach have same quality of the clusters in Figure 1.

Figure 2 describes the Effectiveness of Proposed Algorithm, including Running Time, Number of immune networks and Number of clustering. It is obvious that the higher the number of Spam is, the better the quality of clustering results will be. In the other words the more Spam individual client user feedbacks, the better email serve Spam filters get, the less likely individual client user will be caught. Our algorithm will continuously deliver similar groups of SPAM during the run of algorithm.
In Fig. 2, there are three types of Effectiveness about Proposed Algorithm. (mean value of 50 times) ， (a) describes Running Time changing with number of spam , (b) describes Number of immune networks and Number of clustering changes with number of spam .



(a) **(b)**

Figure **2.** the Effectiveness of Proposed Algorithm. (mean value of 50 times)

The computational cost of our algorithm for clustering crucially depends on the size increment data set of x and size of initial starting dataset. We denote computational cost as follow:

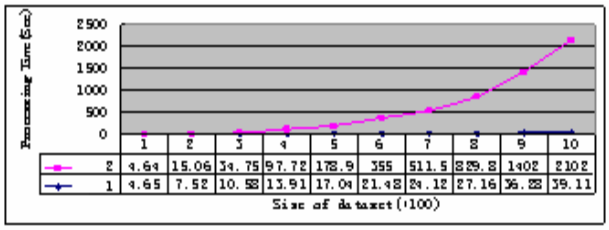


Figure **3.** Comparison of computational cost between incremental algorithm for clustering and re-clustering

In order to maintain clusters in dynamic environment with a high volume of updates without costly re-clustering. We shall see the data stream as a sequence of time ordered “windows” that contain a very limited number of objects at a given time.To sum up ,Here are some of the reasons that we have found our incremental approach to be useful in practice.

5. Summary and Discussion

Spam is a big and complex problem today, so there is a significant financial incentive for Spammers to learn to defeat any Spam-reduction techniques. Because of this reason, any robust, long-term anti-Spam solution must use multiple techniques in several layers, and must involve cooperation among all parties who are interested in finding

solutions. In this paper, being motivated to protect our networks from the escalating costs of Spam, we propose a behavior-based collaborative algorithm for e-mail server using Artificial Immune System, in which the immune-inspired algorithm is capable of continuously identifying similar groups of Spam. There are several interesting directions for future research including compressing practical dataset with complex dynamics and distributions using our incremental algorithm for clustering, we have found our novel approach useful in web dynamic data environment, we have compared the processing time of the incremental algorithm for clustering and re-clustering ,and found that on the average, results indicate that the incremental approach faster than re-clustering ,

However, the new approach has shown its characteristics of reliability, efficiency and scalable ability, which can be used in conjunction with other filtering systems.

References

[1] Ion Androutsopou
los, John Koutsias, Konstantinos V. Chandrinos, Georgios Paliouras, and Constantine D. Spyropoulos(2000). "An Evaluation of Naive Bayesian Anti-Spam Filtering". In Proceedings of the workshop on Machine Learning in the New Information Age,.

[2] Sophos Inc. Field guide to Spam. (2004) http://www.sophos.com/Spaminfo/explained/fieldguide.html. Continuously updated. Last accessed March 2, 2004.

[3] A. Jain and R. Dubes. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, N.J., 1988.

[4] Chen, C., Hwang, S., Oyang, Y. An Incremental Hierarchical Data Clustering Algorithm Based on Gravity Theory. In 6th Pacific Asia Conference on Knowledge Discovery and Data Mining, 2002.

[5] Ester, M., Kriegel, H-P., Sander, J. Wimmer, M., Xu, X. Incremental Clustering for Mining in a Data Warehousing Enviornment. VLDB'98, 323-333, 1998.

[6] Ester, M., Kriegel, H-P., Sander, J., Xu, X. A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD'96, 226-231, 1996.

[7] Widyantoro, D. H., Ioerger, T. R., Yen, J. An Incremental Approach to Building a Cluster Hierarchy. ICDM'02, 705-708, 2002.

[8] Charikar, M., Chekuri, C., Feder, T., Motwani, R.Incremental Clustering and Dynamic Information Retrieval. In 29th Symposium on Theory of Computing, 626-635, 1997.

[9] M. Charikar, C. Chekuri, T. Feder, and R. Motwani, "Incremental clustering and dynamic information retrieval," in Proceedings of STOC,May 1997.

[10] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," In In Proceedings of the AAAI 2000 Workshop on Articial Intelligence for Web Search, pp. 58--64, Austin, Texas, July 2000.

[11]Varela, F. J., Coutinho, A. Dupire, E. & Vaz, N. N. (1988), "Cognitive Networks: Immune, Neural and Otherwise", Theo-retical Immunology, Part II, A. S. Perelson (ed.), pp. 359-375.

[12] de Castro, L. N., et al, (2002) "Artificial Immune System: A New Computational Intelligence Approach", Springer-Verlag.2002

[13] Hofmeyr, S., and S. Forrest(2000), "Architecture for an artificial immune system", Evolutional Computation Journal, vol. 8,no.4, 2000

[14] Jerne, N. K(1974). "Towards a Network Theory of the Immune System", Ann. Immunol. (Inst. asteur), 1974, pp. 373-389.

[15]de Castro, L. N. & Von Zuben, F. J. (2001), "aiNet: An Artificial Immune Network for Data Analysis", in Data Mining: A Heuristic Approach, Hussein A. Abbass, Ruhul A. Sarker, and Charles S. Newton (eds.), Idea Group Publishing, USA, Chapter XII, pp. 231-259.

[16] Justin Mason (2004). "The SpamAssassin Homepage". http://Spamassassin.org/index.html,

Xiaobing Liu, male, 1956.7, Professor in Dalian University of Technology, doctoral supervisor, director of CIMS Center, head of the Institute of Automation Technology of School of Mechanical Engineering, member of the academic committee of Dalian University of Technology and School of Mechanical Engineering as well, Director of Dalian CAD Development and Application Center, initiator of engineering study in Liaoning Province, member of expert committee of Project 863 specialized in CIMS.

Recently, mainly engaged in the research and application of CAD/CAM, CIMS and enterprise information technology and soft computing. A contributor of two monographs and more than 50 articles have been published in core periodicals such as China Mechanical Engineering Journal.

Nan ZHANG, male,1970.3,Senior Engieer，PhD Student of Department of Mechanical Engineering, Dalian University of Technology.

Main Research focus on ERP, MES, enterprise information technology and artificial immune system.