Adaptive E-mail Intention Finding Mechanism based on E-mail Words Social Networks

Che-Fu Yeh[1] Ching-Hao Mao[1] Hahn-Ming Lee[12] Tsuhan Chen[3]
[1]National Taiwan University of Science and Technology, [2]Academia Sinica, [3]Carnegie Mellon University
m9415041@mail.ntust.edu.tw d9415004@mail.ntust.edu.tw hmlee@mail.ntust.edu.tw tsuhan@cmu.edu

ABSTRACT
Through the rapid evaluation of spam, no fully successful solution for filtering spam has been found. However, the spammers still spread spam by using the same intentions such as advertising and phishing. In this investigation, we propose a mechanism of E-mail Words Social Network (EWSN) for profiling users' intentions related to interesting and uninteresting e-mails. An EWSN is constructed from the information in an individual user's mailbox, and expands e-mail information from the World Wide Web (WWW) via the search engine. Based on the web information and association rules among the words, words and relations are expanded as a words' social network. Via the EWSN, both interested and uninterested EWSNs can be constructed to analyze user intentions. Additionally, an efficiency detection mechanism based on the EWSN is proposed to classify e-mails. Finally, the adaptation algorithm of artificial immune system is applied to EWSN, which is thus adapted to follow the user's confirmed classification results. The experimental results indicate that the proposed system is very helpful for classifying spam e-mails by analyzing senders' intentions. Some ideas for analyzing interested nature of people, and profiling their backgrounds, are also presented.

Categories and Subject Descriptors
H.4.3 [Communications Applications]: Electronic mail

General Terms
Design, Security

Keywords
E-mail words social network, social network, artificial immune system, intention finding, spam classification

1. INTRODUCTION
Unsolicited bulk e-mail has become increasingly serious on the Internet recently [1, 2]. Although many anti-spam techniques have been proposed recently [3-16, 20, 23, 35], no fully successful solution for overcoming spam has been found [1]. The heterogeneity, definition and evolution make spam a serious problem. Restated, the rapid evolution of spam approaches and the indistinct criteria of recognizing spam make spam difficult to counter. However, these issues can be resolved by finding senders' intentions, because different techniques and recognition criteria still use the same intentions to spread spam. The spam intentions such as advertising and phishing have the same surface topics such as "click me" or "read me". Therefore, finding the topics and intentions of e-mail senders is an interesting issue [17-22].

Two methods are available for combating spam, i.e., blocking the sources of spam sources [8-13, 15] and filtering spam according to content [3-7, 14, 16, 20, 23, 35]. Spammers can be blocked according to various parameters, such as e-mail addresses and e-mail servers. The rule-based method is employed for this goal. It can identify and block the spam e-mails by using black-lists, white-lists and some heuristic rules [5]. The main disadvantages of rule-based filtering mechanism are that the rules need be maintained manually, and are difficult to evaluate. The sender reputation method is also extensively applied on the web e-mail services. This method can classify e-mails by calculating reputations of users and e-mail servers based on user feedback. Many web e-mail service providers apply reputation to classify sending domains, such as Sender Policy Framework [11], GMAIL [12] and Yahoo Domain Key [13]. One problem is the difficulty of identifying untrustworthy senders, such as senders from zombie networks. In this problem, spammers use viruses and other malicious programs to conscript vast networks of computers belonging to users, without their knowledge, to send spam. Reputation information of senders is very hard to apply to find such spammers. However, the social relation between the legitimate senders and spammers is shown that they have different network structures [8]. Therefore, the social relation between the senders can be inferred from the information of e-mails such as header and content.

Spam can be filtered according to the content of e-mails. Content-based filtering is often applied to generate automatic filtering rules or classifying models via machine learning approaches, such as Naïve Bayes [3] and Support Vector Machine [4]. These approaches usually analyze words and phrases from their appearance and distributions of in the content of e-mails. A mathematical model is then built, and used to filter the incoming e-mails. However, spammers often conceal the content of their e-mails in pictures in order to bypass content-based filters such as image-spam [23]. Therefore, spam classification requires a method of profiling spam effectively.

Social network analysis has emerged as a crucial key method for profiling social relationships [8-10, 25-27, 32, 33, 36, 39, 40].

LSAD'07, August 27, 2007, Kyoto, Japan.

Different social networks represent different attributes of network structures, and can be utilized to quantify the abnormal behavior. In anti-spam research, some social network analysis methods focus on analyzing the information in e-mail headers, such as From, To and CC., and attempt to discover the social relationships from the information [8-10]. Boykin and Roychowdhury [8] proposed a concise method for applying personal e-mail network to filter spam. O'Donnel *et al.* [9] utilized e-mail server log to construct an e-mail social network for detecting unauthorized accounts. Kong *et al.* [10] proposed a collaborative spam filtering based on e-mail social network analysis to improve the efficacy of use of e-mail networks' topological attributes. These methods focus on blocking the spammers, and facilitate information sharing via pervasive social communities in cyberspace. However, they are only appropriate for a general anti-spam solution, rather than for a specified user. Therefore, existing social information can be combined with individual users' intentions to alleviate both the rapid evaluation of spam techniques and the indistinct properties of spam content.

This investigation proposes a mechanism of E-mail Words Social Network (EWSN) for profiling users' intentions related to interesting and uninteresting e-mails. The processes of constructing EWSN are also described. At first, the proposed method expands the words of E-mail content by mining words information from Internet, forming the EWSN. Then by applying the Social Network Analysis (SNA), our method can easily distinguish non-spam e-mails and spam e-mails through the network properties of different E-mail contents. At last, a modified Resource-limited Artificial Immune Network algorithm (RAIN) [24] in artificial immune network (AIS) is adopted for adapting user's feedback on mail classification. Through the EWSN, both interested and uninterested EWSNs can be constructed to analyze user intentions. The experimental results indicate that analyzing senders' intentions is suitable for classifying e-mails.

Section 2 introduces the social network of words, and describes in detail the EWSN, which is proposed in this investigation to classify spam. Section 3 first introduces the system overview of EWSN. Section 4 describes the experiments performed on the proposed systems. Section 5 draws conclusions.

2. E-MAIL WORDS SOCIAL NETWORK

This section first introduces the social network of words [25, 26]. The structure and properties of e-mail words are then introduced and discussed. Finally, an e-mail words social network is introduced to profiling users' interest or lack of interest.

2.1 Words social network

A Words Social Network (WSN) gives a possible explanation for some coincidence. For example, two different people give the same comments when they see the specified pictures. WSN might show an idea about analyzing association memory of human [26]. WSN also reveals high clustering, which allows searching by association. Constructing social networks by words satisfy the power law as long as they agree with scale-free network [26, 27]. Motter *et al.* [26] described a network constructed from words in an encyclopedia. They observed that the node-degree distribution follows a power law, and that most nodes are linked to few other nodes. Ferreira *et al.* [25] found that a WSN is clustered according to either events or persons. They devised a questionnaire for two social groups. The questionnaire comprised three questions, each for a different topic. The author recorded the results, and created WSNs when they got responses with synonyms about a topic. Some different feedback keywords were found in the results of the two social groups, but the most of words in the WSN were the same. This finding indicates that the topics and intentions have more effects on feedback keywords than different social groups. Therefore, the WSNs can discover the association words of a topic. Their survey results inspire us that WSN can reveal the intentions and thoughts of people.

2.2 The structure of E-mail words

The words of e-mail content depend on users' interest and writing intentions. However, spam and non-spam e-mails cannot easily be distinguished without considering users' interests. People might define spam e-mails by their different interests. Secker *et al.* [6, 7] differentiate e-mail as "interested and uninterested" rather than "spam or non-spam" from the mailbox. This idea derives from the basic artificial immune system [28-30], where the concept is "self and non-self", and regards non-spam e-mail as interested e-mail, and spam e-mail as uninterested e-mail. Their result demonstrates that two structures in e-mail can be discriminated by different words corresponding to interested and uninterested. Because e-mail can be classified into two structures, WSN can be applied to extract the topics determining whether emails are interested or uninterested. Li and Hsieh [15] noted that similar groups of spammers send similar e-mails. This similarity increases the possibility of finding similar writing intentions or writing purposes. Using this approach, this investigation presents an E-mail Words Social Network (EWSN) that can profile the intentions of various e-mail structures.

2.3 E-mail words social network

EWSN attempts to extract profiles of user intentions from content information given by the e-mail inbox. Compared with WSN, the proposed EWSN has three different features. The first feature is the linking method. Through the research of words' social network [25, 26], the link between each node is defined by synonyms. Figure 1 illustrates synonyms defined methods such as dictionary, encyclopedia [26] and oral definition [25]. In EWSN, synonyms are defined from the specific words in e-mail headers. For instances, three words, "mouth", "hygiene" and "shower", are extracted from an e-mail's subject. These words could be synonyms based on the topic "clean". The second feature is a web mining method [31-34] that mines web information for expanding social relations in e-mails. Martin-Bautista *et al.* [31] proposed a system to solve query expansion problems, using a fuzzy association rule to measure the associations among words for query expansion. Based on the association rule and query results, words and relations can be employed to construct the EWSN. The last feature is Resource-limited Artificial Immune Network algorithm (RAIN) [24], which is modified to adapt EWSN. The RAIN allocates the network size, and manages stimulation on each node. Because the nodes in EWSN are grouped by linking synonyms, the stimulation method of nodes is changed into groups, and added to RAIN.

An EWSN can profile interested and uninterested nature of a user through the attributes of WSN. Conversely, web information not only enriches social relationships in EWSN, but also updates

words associated with spam regularly. These updated spam words allow EWSN to catch the trend of spam. The user can check the classifying results and respond misclassification when testing EWSN using a testing set. The modified RAIN adopts the received feedback of users to adapt the EWSN. Accordingly, the EWSN can be applied to profile intentions of e-mailers. Furthermore, the EWSN can analyze interested nature of people to profile their backgrounds.

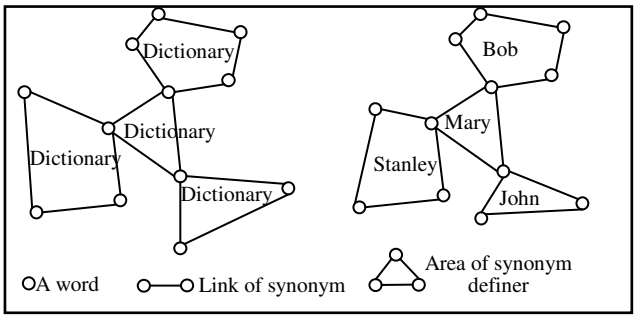

Figure 1: Words social network.

3. THE SYSTEM ARCHITECTURE OF EWSN

This section presents the system architecture of EWSN, as shown in Figure 2. The system is comprised of three modules, namely the Word Social Network Constructor, the Concept-based Detector and the Immune-based Network Adaptation. Additionally, the system utilizes three data sources, namely Labeled mail data, Web information and Background knowledge. These modules and data sources will be briefly described as follows. The Word Social Network Constructor is responsible for expanding the words of e-mail and constructing the words' social network. The network can be applied to represent concepts of e-mail once it is constructed. The Concept-based Detector is used to detect the similarities and the growth rate of concepts. The Immune-based Network Adaptation is used to adapt EWSN by combining user feedback. Labeled mail data adopts the Spam-Assassin dataset to train the model. It contains spam and non-spam, where the set of non-spam is also divided into hard-non-spam and easy-non-spam by hand-classified. Web information is obtained from the Internet, which is considered as a large database for expanding the novel and high relational word related to the words in e-mails. The Google search engine can be used to expand the list of e-mail words by querying the e-mail words**.** Background knowledge refers to the application of some dictionary and heuristic knowledge as the background database for reducing the noise words.

Additionally, EWSN has three phases, namely the Training, Testing and Adapting phases. These phases consist of different modules, depending on their aims. The main role of the training phase is to build the E-mail Words Social Network Constructor, which is used to construct the word concept network. It can integrate the mail data labeled by users, and automatically extends words from web information. Additionally, it adopts background knowledge to remove the noise words from the e-mail. The profiled concepts of spam and non-spam e-mails are recorded in the E-mail Words Social Network database. In the testing phase, the Concept-based Detector is used to classify newly arrived e-mail messages as spam or non-spam based on the EWSN generated from the training phases. Finally, in the adapting phase, the EWSN is continually adjusted. The Immune-based Conceptual adaptor can be used to handle the adaptation on the words network, and modifies the EWSN database based on user response. The mechanism of each module is described in detail below.

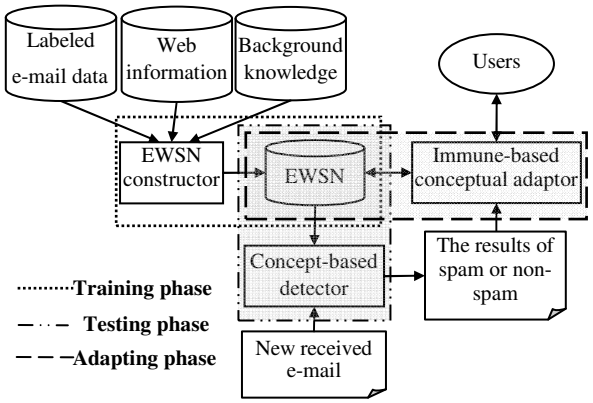

Figure 2: The system architecture of e-mail words social network.

3.1 E-mail words social networks constructor

The EWSN Constructor can profile the sender mailing intentions from labeled e-mail, and generates both spam and non-spam EWSNs. This module has four components, namely Novel Word Extension, Word Association, Word Relational Linking and Writing Intention Labeling, as shown in Figure 3. Because an EWSN is built from words, each labeled mail data needs to be preprocessed first.

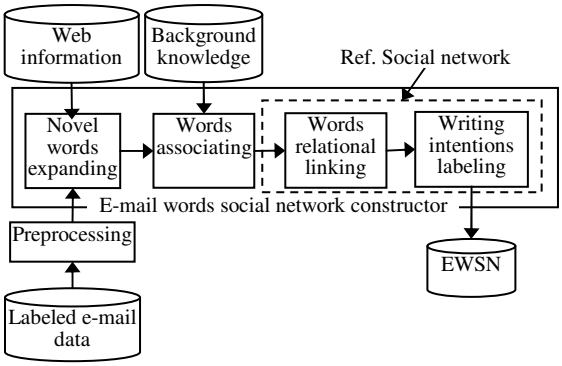

Figure 3: The architecture of e-mail words social networks constructor.

3.1.1 Preprocessing unit

A preprocessing unit is used to separate e-mail messages into words as nodes of EWSN. It is based on two nature language processing approaches, stop-word filtering and word stemming of spam [14] and the concepts come from nature language processing. E-mail messages are separated by the "Subject" and "From" headers to reduce noise of words [6, 16]. Secker *et al.* [6] proposed a data coding that uses data from various e-mail headers, such as "From", "To" and "Subject", because these data are simple to encode for training. Noise words are removed from the e-mail messages using stop-word filtering, and the basic, or uninflected, form of each word is obtained using word stemming

of spam [14]. Through our observation, the most of remaining normal words after removing stop-words and inflections are nouns and verbs, which can be used to define synonyms. In the proposed approach, every e-mail input module needs to perform preprocessing.

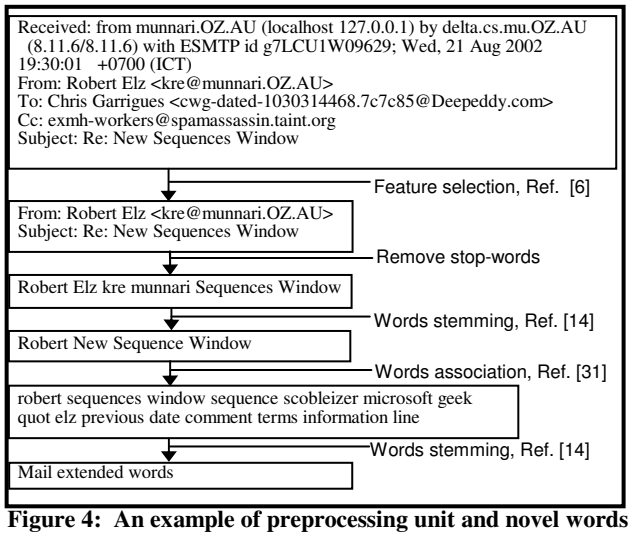

Figure 4: An example of preprocessing unit and novel words expansion.

3.1.2 Novel words expanding

This component separates the e-mail into spam and non-spam by following the labels, and transforming them into a word group to discover novel words from the web. The sliding window is used to slice the e-mail messages to a word group, which is entered as a query into a web search engine. The results of novel words are obtained from the web search engine, and are selected according to query length and exponential operations. For example, if the query consists of two words, then only the first four results are chosen, because 2^2=4. However, if the query consists of ten words, then 1024 results have to be chosen. Therefore, an upper bound is set on the number of results received. Figure 4 illustrates an example of preprocessing units, and novel words expansion.

3.1.3 Words associating

When receiving novel words, the TFIDF [35] and association rule [31] are applied to identify strongly related novel words. Martin-Bautista *et al.*[31] presented the fuzzy association rule to calculate the relationships among words. These related novel words are used to provide a web perspective synonym, enabling e-mails to be extended through the web perspective. In this component, the knowledge base database is also applied to identify the words that belong to an idiom. Because an idiom is a general term in writing, its effect in EWSN needs to be lowered. Therefore, the weightings of words matching the idiom database were decreased. The formula of association rule is shown in Equation (1)-(3):

$$p(t_{ij}) = \frac{\left|\left\{T \in S \,\middle|\, t_{ij} \subseteq T\right\}\right|}{|S|} \tag{1}$$

Where $T = \{t_1 \ldots t_n\}$ is a collection of extracted word sets from querying results. $S = \{s_1 \ldots s_n\}$ is a collection of feedback results by search engine, and the text content of each feedback result s_k is used to obtain a set of keywords $t_k = \{t_{k1} \ldots t_{km}\}$. The support of the word t_{ij} defined as the probability of finding t_{ij} in a search engine feedback result of S.

$$Supp(t_{ij} \rightarrow t_{ij+1}) = p(t_{ij} \cup t_{ij+1}) \tag{2}$$

$$Conf(t_{ij} \rightarrow t_{ij+1}) = \frac{p(t_{ij} \cup t_{ij+1})}{p(t_{ij})} \tag{3}$$

Where the support and confidence of the rule $t_{ij} \rightarrow t_{ij+1}$ noted by $Supp(t_{ij} \rightarrow t_{ij+1})$ and $Conf(t_{ij} \rightarrow t_{ij+1})$, respectively.

3.1.4 Words relational linking

To build EWSN, synonyms are applied to link words output from the word association module. Because the synonyms are defined from e-mails, each e-mail represents a clique in EWSN. These synonym links generate a graph, which can be applied for social network analysis [36, 39, 40].

3.1.5 Writing intentions labeling

In this part, betweenness and cliques are calculated according to social network analysis methods [36, 39, 40]. The betweenness represents the number of times a node is passed through. The variation of in-degree and out-degree varies according to betweenness. Therefore, a node with high betweenness indicates the sender's writing intentions. For example, a high betweenness means that senders who have the same written content have good possibility of the similar writing intentions. A way of visualizing the model is still required to identify writing intentions. A clique can be represented as an e-mail, because the synonyms are defined with e-mail. Calculating cliques also helps to observe the variation inside a network. The cliques in EWSN reveal the number of e-mails defined within it. Additionally, the growth rate of cliques shows the linking status of EWSN when expanding novel words.

3.2 Concept-based detector

This module has two components, namely Concept stimulation measuring and Concept variant rate testing, as shown in Figure 5. In the proposed approach, the words are extracted from e-mail, and then applied as query input to the search engine to retrieve related words. Social network analysis is employed to discover the relative cliques' density of two e-mails from different sources. Because the synonyms are defined in terms of e-mail, each e-mail is represented in a clique.

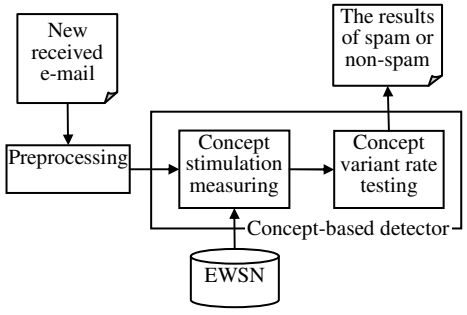

Figure 5: The architecture of concept-based detector.

3.2.1 Concept stimulation measuring

Concept stimulation measuring involves first calculating each node's centrality degree and centrality betweenness, as described in social network analysis [36, 39, 40]. The centrality degree measures the in-degree and out-degree of a node. Centrality betweenness measures the number of times that a node has passed between other two nodes on the shortest path. Network computer programs, such as UCINET [36], apply these two measurements. Each node has a weight value following measurement. The weight can then be used for stimulation measurement. The stimulation calculates the weighted sum of mapping input words.

The formula of centrality betweenness is shown in Equation (4), for a graph $G = (V, E)$, G consists of a finite nonempty set of vertices V, and a finite set of edges E. The Centrality betweenness of vertex n is defined as follows:

$$C_B(n) = \sum_{j<k} \frac{Count(g_{jk}, n)}{g_{jk}}, \; n \neq j \neq k \in V \qquad (4)$$

Where g_{jk} is defined as the number of geodesic paths between vertex j and vertex k, and $Count\ (g_{jk}, n)$ be the number of these geodesics pass through vertex n.

3.2.2 Concept variant rate testing

In concept variant rate testing, the growth rate of cliques in the network is measured. The high weight words of input e-mail are merged into the EWSN. Specifically, words that do not exist in EWSN and have high betweenness are added to the EWSN. Because the words belonging to a clique have weak relationships, the words obtained from the network are more likely to have weak relationships than highly related words. Such a clique is called a weak clique in social network analysis. A hard e-mail has low stimulation in the EWSN. Therefore, only the hard e-mails are chosen for merging, because the time taken in this component has to be decreased. Finally, the module outputs the classification results of the e-mail.

3.3 Immune-based conceptual adaptor

Figure 6 illustrates Immune-based conceptual adaptor module, which has four components namely misclassified checking, word adapting, stimulation word weighting and stimulation concept weighting. The proposed adaptive method is based on artificial immune networks [24, 28-30]. First, in the misclassified checking component, user feedback is combined with our adaptive algorithm, in order to help update the user's interest and adapt the EWSN. Then word adapting and stimulation word weighting are then performed, using Resource-limited Artificial Immune Networks (RAIN) [24], to adapt the model automatically. RAIN simulates all nodes on the network with each other. For example, if a network node A detects a pathogen, then its weight is increased, and those of other nodes, which can discriminate node A, will decrease their own weights. The aim of this procedure is to find nodes that can detect the most unknown pathogen. The last component in this module modifies the RAIN algorithm to fit EWSN. In stimulation concepts weighting, our concepts are defined as cliques in EWSN. Because the proposed approach calculates the growth rate of cliques, RAIN is also applied to modify the weights of words in cliques. Using the above procedure, if the user reports some words as being correctly classified, then modified RAIN is applied to emphasize the words and related groups. Otherwise, the weight of the misclassified words is decreased. The modified RAIN receives feedback, and uses it to adapt EWSN, every time the users respond misclassification.

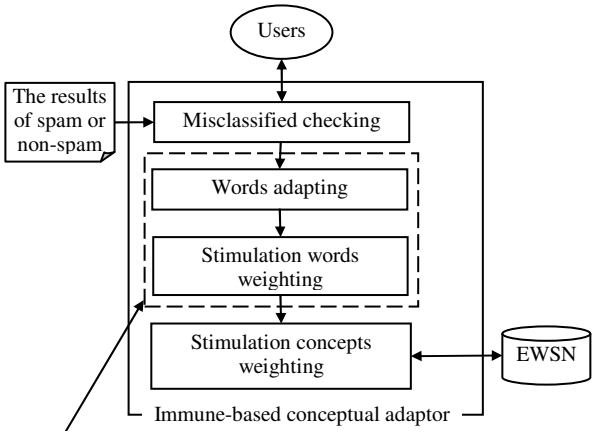

Figure 6: The architecture of immune-based conceptual adaptor.

4. EXPERIMENTAL RESULTS

This section presents the experiment results of the proposed method. Two experiments were performed to analyze EWSN. The first experiment was to analyze the effect of different inputs on EWSN. The second experiment compared EWSN with Support Vector Machine (SVM) [4] and Naïve Bayes (NB) [3], and recorded accuracy, precision and recall. The dataset and evaluation methods used herein are introduced. The experimental procedures and results are then discussed.

4.1 Dataset

The SpamAssassin (SA) corpus [42] and the Text REtrieval Conference (TREC) 2005 public spam corpus [43, 44] were used for our experiments. The SA dataset contains non-spam (legitimate e-mail) and spam e-mails collected from the SA developer mailing list. The non-spam also comprises two sets, called easy-non-spam and hard-non-spam. This dataset is used popularly in evaluations of publicly available spam filters.

The TREC 2005 spam corpus contains 92189 messages which consist of 52790 spam e-mails and 39399 ham e-mails. This dataset also provide four subset index labels. The dataset is also extensively used to evaluate spam filtering approaches.

Our experiments adopted only two e-mail header fields "Subject" and "From" following [6, 16]. The corpora were preprocessed at the beginning of the experiment by extracting these fields. We built three testing sets T1, T2 and T3 to simulate unbalance dataset [41]. T1 uses the messages from the SA corpus [42], and T2 and T3 use the messages from the TREC 2005 spam corpus [43, 44]. A preliminary testing set T1 was generated by mixing 100 randomly chosen e-mails from non-spam, and 1300 e-mails from spam. The testing set T2 was generated by mixing 1250 randomly chosen e-mails from non-spam, and 5000 e-mails from spam. The last testing set T3 uses 5000 randomly chosen non-spam e-mails, and 1250 randomly chosen spam e-mails.

4.2 Evaluation and tools

A confusion matrix [38] was applied to measure accuracy, recall and precision in our experiments. A confusion matrix contains information about actual and classified result obtained by a classification system. The spam classification result can be evaluated by using the accuracy and recall measure. To compare with other methods, the open source WEKA [37] framework was employed to test the classifiers. WEKA is a popular Machine Learning (ML) tool, which includes many feature selection and training algorithms. It is well-suited for developing new data mining systems. The UCINET [36] software was also employed to visualize our EWSN. UCINET is a collection of social network algorithms for social network analysis.

4.3 Experiment 1

The aims of Experiment 1 were to determine whether an EWSN built by different amount of e-mails works well, and whether different e-mail classifications affect EWSN. The results and discussions are given below:

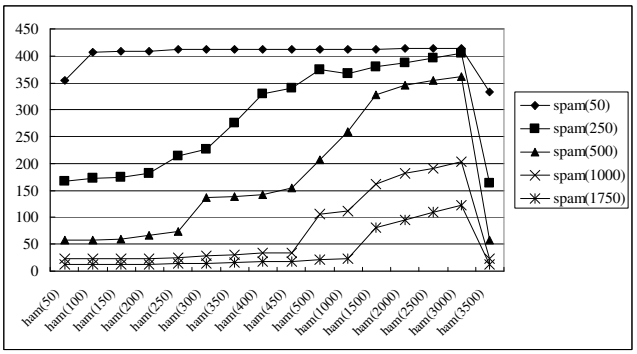

Figure 7: The amount of mis-classified non-spam e-mails.

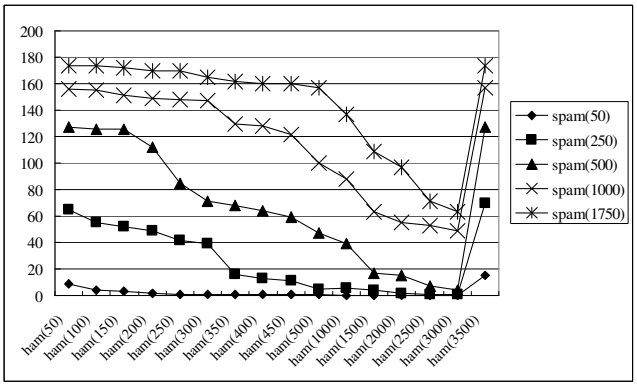

Figure 8: The amount of mis-classified spam e-mails.

4.3.1 The effect of different inputs

We use the SA corpus to make 10-fold cross validation and analyze the classified influence from the amount of e-mails. The training e-mails are 0 to 1750 spam e-mails and 0 to 3500 non-spam (ham) e-mails. Both of Figure 7 and Figure 8 show that the amount of training e-mails are positive interrelated with error rate. In addition, the accuracy of two networks in EWSN are significant affected by the noises of them. Further, spam and non-spam were input individually to build the EWSN, and UCINET was employed to visualize the EWSN. Two interesting structures were observed in the EWSN. Figure 9 illustrates one finding. The left of Figure 9 displays three closure networks with no overlapping. The right of the picture displays three networks expanded from the network on the left-hand side. Most of the structures in Figure 9 are built from spam. Figure 10 illustrates structures consisting of mostly of non-spam e-mails. Moreover, the results can be applied to verify other variants inside EWSN.

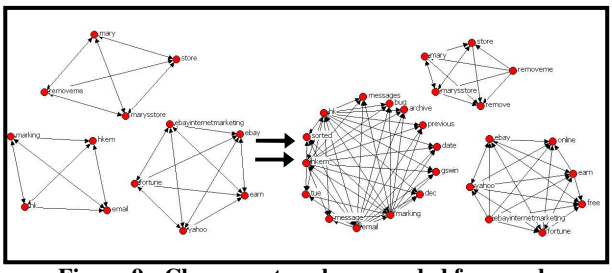

Figure 9: Closure networks expanded from web.

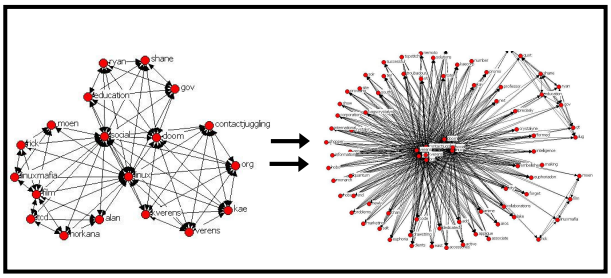

Figure 10: Intersection networks expanded from web.

4.3.2 The growth rate of cliques

Throughout the preliminary analysis, words were input to EWSN one by one, and the social network analysis method from UCINET was employed to check variants inside. The growth rate of cliques indicates significant differences between words belonging to spam and words belonging to non-spam. The cliques in EWSN are defined by synonyms. Due to overlapping of words, 50 e-mails were input to EWSN, so that it contained at least 50 cliques. In this experiment, the EWSN had 276 cliques after training 50 spam e-mails, and 888 cliques after training 50 non-spam e-mails. Experimental results reveal that spam and non-spam create significantly different EWSN cliques. The growth rate of cliques is high when EWSN expands input-words of non-spam. These words are often names, organization or specific terms. Conversely, EWSN built by spam produces cliques with low growth rates, because spam words have strong relations in the Web. In our observation, spam words have strong relations in the web because they are frequently used in the same concepts or topics. Based on this phenomenon, continuing work will be to enhance EWSN in increase the speed of checking.

4.4 Experiment 2

In Experiment 2, EWSN was compared with SVM and NB. We trained two models and used WEKA to obtain the results of SVM and NB. The first training was performed with 100 e-mails comprising 50 spam and 50 non-spam, and the second training was performed 600 e-mails comprising 300 spam and 300 non-spam. Table 1 lists the results over the testing set T1. EWSN produced a higher accuracy than other two methods in small training set, because the number of features was too small. The SVM was found to have very good accuracy, precision and recall when 1000 e-mails were input to train SVM and NB. However,

the number of spam e-mails is much greater than the number of non-spam e-mails in the real world. Due to this imbalance, Figure 11 and Figure 12 illustrate that EWSN is more effective at overcoming this problem of small training examples than other methods. Additionally, the non-spam (ham) recall and precision must be considered in order to avoid losing non-spam emails. The results of calculating the *f*-measure in non-spam (ham) from Table 1 shows that EWSN has a better result, of 0.209, than SVM (0.177) and NB (0.14). In small training set, EWSN still has to enhance its non-spam-recall and precision.

Table 1: Experiments comparison with two major methods

Training method (input examples)	Accuracy	Spam-precision	Spam-recall	Ham-precision	Ham-recall
SVM (100 e-mails)	37.58%	98.6%	33.2%	9.8%	94%
NB (100 e-mails)	12.03%	99%	5.2%	7.5%	99%
EWSN (100 e-mails)	87.72%	94.34%	86.00%	15.28%	33%
SVM (**1000** e-mails)	99.72%	99%	99.8%	96.9%	99%
NB (**1000** e-mails)	98.99%	99.8%	99.1%	89.1%	98%
T1 = 1400 e-mails (1300 spam, 100 non-spam)					

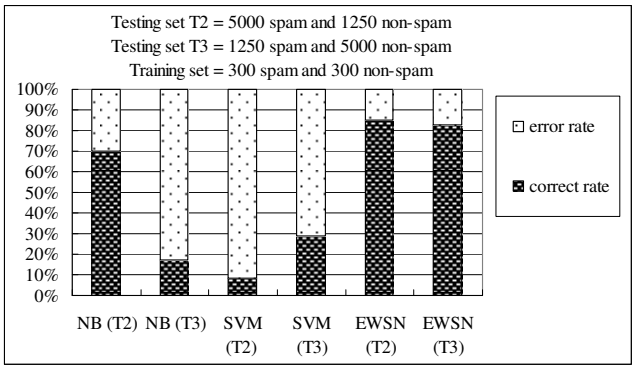

Figure 11: The accuracy comparison of three methods.

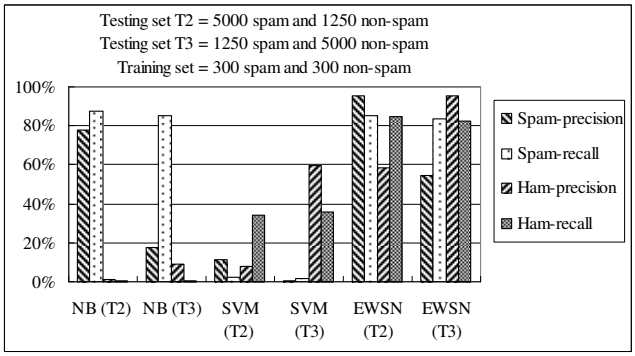

Figure 12: The precision and recall comparison of three methods.

5. CONCLUSION AND FURTHER WORK

In this investigation, we propose an Email Word Social Network (EWSN) for finding senders' intentions. EWSN is inspired by social networks and Artificial Immune Systems (AIS). The concept of social networks is applied to build an EWSN, which is adapted by AIS. EWSN yields some interested results in e-mails. Experiments 1-2 reveal several interesting issues:

Clique growth rate: The use of different classification of e-mails to build EWSN reveals that cliques have different growth rates. This property can be exploited to enhance recall rate of non-spam by nearly 40% upgrading.

Unbalance dataset suitability: EWSN can reduce the imbalance found in many datasets by using information from the Web. However, this approach raises system overhead due to the need to remove noise from training data. A method to decrease system overhead is thus required. Otherwise, because web information changes very rapidly, this method is dependent on search engines.

Extracting intentions from e-mail is an example for profile information. Analyzing interested nature of people can be used to profile their backgrounds. Further work will test EWSN with other corpora and try to create their background information. This investigation can hopefully contribute towards fighting spam, and combining AIS with social network concepts to profile people.

6. ACKNOWLEDGMENT

This work was supported in part by the National Science Council of Taiwan under grants NSC 95-2221-E-011-150, NSC 95-2221-E-011-151, moreover by the Taiwan Information Security Center (TWISC), the National Science Council under grants NSC 95-2218-E-001-001, and NSC 95-2218-E-011-015, and also by the iCAST project, the National Science Council of Taiwan under grant NSC95-3114-P-001-002-Y02.

7. REFERENCES

[1] L. H. Gomes, C. Cazita, J. M. Almeida, V. Almeida, and W. M. Junior. Workload models of spam and legitimate e-mails. *Performance Evaluation*, 64(7-8):690-714, August 2007.

[2] A. J. Donnell. The Evolutionary Microcosm of Stock Spam Oapos.*IEEE Security & Privacy Magazine*, 5(1):70-75, 2007.

[3] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk e-mail. *AAAI Workshop on Learning for Text Categorization*, pages 55-62. 1998.

[4] H. D. Drucker, D. Wu, and V. Vapnik. Support Vector Machines for spam categorization. *IEEE Trans. on Neural Networks*, 10(5):1048-1054, 1999.

[5] J. G. Hidalgo and M. M. Lopez. Combining text and heuristics for cost-sensitive spam filtering. *Computational Natural Language Learning Workshop*, pages 99-102. 2000.

[6] A. Secker, A. A. Freitas, and J. Timmis. AISEC: An Artificial Immune System for Email Classification. *The IEEE Congress on Evolutionary Computation Proceedings*, 1: 131-138, December 2003.

[7] T. Oda and T. White. Immunity from Spam: An Analysis of an Artificial Immune System for Junk Email Detection. *The 4th International Conference on Artificial Immune Systems*, pages 276-289. August 2005.

[8] P. O. Boykin and V. P. Roychowdhury. Leveraging social networks to fight spam. *Computer*, 38(4):61–68, April 2005.

[9] A. J. O'Donnell, W. C. Mankowski, and J. Abrahason. Using E-mail Social Network Analysis for Detecting Unauthorized Accounts. In *Proceedings of the Third Conference on Email and Anti-spam*, July 2006.

[10] J. S. Kong, B. A. Rezaei, N. Sarshar, V. P. Roychowdhury, and P. O. Boykin. Collaborative Spam Filtering Using Email Networks. *Computer*, 39(8):67-73, August 2006.

[11] M. Wong and W. Schlitt. Sender Policy Framework (SPF) for Authorizing Use of Domains in E-mail, Available at: http://www.openspf.org/Project_Overview.

[12] B. Taylor. Sender Reputation in a Large Webmail Service. In *Proceedings of the Third Conference on Email and Anti-spam*, July 2006.

[13] DomainKeys, Proving and Protecting Email Sender Identity, Available at: http://antispam.yahoo.com/domainkeys.

[14] S. Ahmed and F. Mithun. Word Stemming to Enhance Spam Filtering. In *Proceedings of the First Conference on Email and Anti-Spam*, July 2004.

[15] F. li, and M. H. Hsieh. An Empirical Study of Clustering Behavior of Spammers and Group-based Anti-Spam Strategies. In *Proceedings of the Third Conference on Email and Anti-spam*, July 2006.

[16] J. Goodman, and W. T. Yih. Online Discriminative Spam Filter Training. In *Proceedings of the Third Conference on Email and Anti-spam*, July 2006.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

[18] V. H. Tuulos and H. Tirri. Combining Topic Models and Social Networks for Chat Data Mining. *IEEE/WIC/ACM International Conference on Web Intelligence Proceedings*, pages 206- 213. September 2004.

[19] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and Role Discovery in Social Networks. *International Joint Conference on Artificial Intelligence*, August 2005.

[20] A. C. Surendran, J. C. Platt, and E. Renshaw. Automatic Discovery of Personal Topics to Organize Email. In *Proceedings of the Second Conference on Email and Anti-Spam*, July 2005.

[21] D. M. Blei and J. D. Lafferty. Correlated topic models. *Advances in Neural Information Processing Systems*, 18:147-154, 2006.

[22] M. W. Berry. *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer-Verlag. 2003.

[23] G. Fumera, I. Pillai and F. Roli. Spam Filtering Based On The Analysis Of Text Information Embedded Into Images. *Journal of Machine Learning Research*, 7:2699-2720, December 2006.

[24] L. N. De Castro, and J. Timmis. *Artificial Immune Systems: A New Computational Intelligence Approach.* Springer-Verlag. London, September 2002.

[25] A. A. A. Ferreira, G. Corso, G. Piuvezam, and M. S. C. F. Alves. A Scale-Free Network of EvokedWords. *Brazilian Journal of Physics*, 36(3A), September 2006.

[26] A. E. Motter, A. P. S. de Moura, Y. C. Lai, and P. Dasgupta. Topology of the conceptual social network of language. *Physical Review E*, 65, June 2002.

[27] H. Ebel, L. I. Mielsch, and S. Bornholdt. Scale-free topology of email networks. *Physical Review E*, 66, 2002.

[28] S. A. Hofmeyr and S. Forrest. Immunity by Design: An artificial Immune System. *Genetic and Evolutionary Computation Conference*, 1999.

[29] S. A. Hofmeyr and S. Forrest. Architecture for an artificial immune system. *Evolutionary Computation journal*, 8(4):443-473, 2000.

[30] P. S. Andrews and J. Timmis. Inspiration for the next generation of artificial immune systems. *International Conference on Artificial Immune Systems*, pages 126-138. 2005.

[31] M. J. Martin-Bautista, D. Sanchez, J. Chamorro-Martinez, J. M. Serrano, and M.A. Vila. Mining web documents to find additional query terms using fuzzy association rules. *Fuzzy Sets and Systems*, 148(1):85-104, November 2004.

[32] A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the Web. In *Proceedings of the First Conference on Email and Anti-Spam*, July 2004.

[33] R. Bekkerman, A. McCallum. Disambiguating Web Appearances of People in a Social Network. *International World Wide Web Conference*, pages 463-470. May 2005.

[34] D. Shen, J. T. Sun, Q. Yang, Z. Chen. Building Bridges for Web Query Classification. *The 29th ACM International Conference on Research and Development in Information Retrieval*, pages 131-138. August 2006.

[35] B. S. Richard and O. K. Jeffrey. MailCat: an intelligent assistant for organizing e-mail. *The third annual conference on Autonomous Agents*, pages: 276 - 282. 1999.

[36] UCINET, The Software for Social Network Analysis, Available at: http://www.analytictech.com/downloaduc6.htm

[37] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[38] R. Kohavi and F. Provost. Glossary of terms. *Machine Learning*, 30:271-274, 1998.

[39] S. Wasserman, and K. Faust. *Social Networks Analysis: Methods and Applications.* Cambridge University Press. 1994.

[40] P. J. Carrington, J. Scott, and S. Wasserman. *Models and Methods in Social Network Analysis.* Cambridge University Press. 2005.

[41] T. Fawcett. "In vivo" spam filtering: A challenge problem for data mining. *KDD Explorations*, 5(2):140-148, December 2003.

[42] SPAMASSASSIN, The SpamAssassin corpus, Available at: http://spamassassin.apache.org/publiccorpus/.

[43] G. V. Cormack and T. R. Lynam. TREC 2005 Spam Track Overview. *Fourteenth Text REtrieval Conference*, 2005.

[44] G. V. Cormack and T. R. Lynam. Spam corpus creation for TREC. In *Proceedings of the Second Conference on Email and Anti-Spam*, July 2005.