2-4 Clustering and Feature Selection Methods for Analyzing Spam Based Attacks

SONG Jungsuk

In recent years, the number of spam emails has been dramatically increasing and spam is recognized as a serious internet threat. Most recent spam emails are being sent by bots which often operate with others in the form of a botnet, and skillful spammers try to conceal their activities from spam analyzers and spam detection technology. In addition, most spam messages contain URLs that lure spam receivers to malicious Web servers for the purpose of carrying out various cyber attacks such as malware infection, phishing attacks, etc. In order to cope with spam based attacks, there have been many efforts made towards the clustering of spam emails based on similarities between them. The spam clusters obtained from the clustering of spam emails can be used to identify the infrastructure of spam sending systems and malicious Web servers, and how they are grouped and correlate with each other. Therefore, we propose an optimized spam clustering method, called O-means, based on the K-means clustering method, which is one of the most widely used clustering methods. By examining three weeks of spam gathered in our SMTP server, we observed that the accuracy of the O-means clustering method is about 87% which is superior to the previous clustering methods. In addition, we define new 12 statistical features to compare similarities between spam emails, and we propose a feature selection method to identify a set of optimized features which makes the O-means clustering method more effective. With our method, we identified 4 significant features which yielded a clustering accuracy of 86.33% with low time complexity.

Keywords
Spam, Clustering, Feature selection, Botnet, Malicious URLs

1 Introduction

In recent years, the number of spam emails has been dramatically increasing and spam is recognized as a serious internet threat. Most recent spam emails are being sent by bots which often operate with others in the form of a botnet, and skillful spammers try to conceal their activities from spam analyzers and spam detection technology. For example, botnet owners control their bots carefully so as to send an extremely small amount of spam from each bot, so that they are able to avoid traditional bot detection technology which checks a quantity of traffic sent by remote systems. In fact, most recent bots are used for sending only 1–2 spam messages on average[1] and in our experiments, we confirmed that each spam sending system sent 1.9 emails on average. Furthermore, the active time of bots and the effective lifetime of a single spam message is highly short. By some estimates, 75% of bots are active for just 2 minutes or less and the lifetime of 65% spam messages is 2 hours or less[2][3].

On the other hand, most spam messages also contain URLs that lure spam receivers to malicious Web servers for the purpose of carrying out various cyber attacks such as malware infection, phishing attacks, etc. Thus, we also need to analyze Web pages linked to URLs in order to determine whether a Web

page is malicious or not. In many cases, however, it is very difficult to analyze all Web pages in real-time, because more than 90% of all email today is considered spam[4], and is abused for various purposes. To make matters worse, real malicious Web servers appear beyond 4 or more Web pages from the initial one linked to URLs directly[5][6]. Therefore, we also need to reduce the analysis time of spam based attacks.

In order to cope with spam based attacks, there have been many efforts made towards the clustering of spam emails based on similarities between them[1][7]-[10]. From the clustering of spam emails, we are able to categorize spam into clusters based on shared similarities. By analyzing each spam cluster, we are able to identify the infrastructure of the systems used for sending spam emails and how they are grouped with each other, and identify the correlation between spam sending systems and malicious Web servers, because in spam based attacks, attackers have to prepare both of them for carrying out their attacks successfully. In addition, if we use cluster information in the classification of spam, it can be used to minimize the time needed for analyzing Web pages. In other words, if an email belongs to a certain cluster, then we do not have to follow its URLs and analyze their Web pages anymore. Therefore, it could be said that the clustering of spam emails is essential to analyze spam based attacks effectively, and also it is very important to improve the accuracy of the spam clustering as much as possible so as to analyze spam based attacks more accurately.

We present an optimized spam clustering method, called O-means, based on the K-means clustering method[11], which is one of the most widely used clustering methods. The K-means clustering method partitions a set of data into k clusters through the following steps.

- Initialization: Randomly choose k instances from data set and make them initial cluster centers.
- Assignment: Assign each instance to the closest center.
- Updating: Replace every cluster's center with the mean of its members.
- Iteration: Repeat Assignment and Updating until there is no change for each cluster, or other convergence criterion is met.

The popularity of the K-means clustering method is largely due to its low time complexity, simplicity and fast convergence. However, it has been well known that its clustering results heavily depend on the chosen k initial centers, and it is very difficult to predefine the proper number of clusters, i.e., k. The O-means clustering method improves its performance by overcoming the shortcomings of the K-means clustering method. By examining three weeks of spam gathered in our SMTP server, we observed that the accuracy of the O-means clustering method is about 87% which is superior to the previous clustering methods. In addition, we define new 12 statistical features to compare similarities between spam emails, and we propose a feature selection method to identify a set of optimized features which makes the O-means clustering method more effective. With our method, we identified 4 significant features which yielded a clustering accuracy of 86.33% with low time complexity.

The rest of the paper is organized as follows. In Chapter **2**, we give brief description for the previous spam clustering methods. In Chapter **3**, we present the proposed clustering method, and experimental results and discussion are given in Chapters **4** and **5**, respectively. Finally, we present concluding remarks and suggestions for future study in Chapter **6**.

2 Related work

Zhuang et al. developed new techniques to map botnet membership using traces of spam emails[8]. Their clustering method is based on the assumption that spam emails with similar content are often sent from the same spammer or attacker, because these email messages share a common economic interest. They identified hundreds of botnets by grouping similar spam messages and related spam campaigns. Li et al. investigated the clustering structures of spammers based on spam traffic collected

at a domain mail server they constructed[7]. In this approach, they used URLs in spam emails as the criterion for the clustering, and they observed that the relationship among the spammers has demonstrated highly clustering structures. In [1], Xie et al. focused on characterizing spamming botnets. To this end, they applied polymorphic URLs which have the same domain name to grouping spam emails. They identified 7,721 botnet-based spam campaigns together with 340,050 unique botnet host IP addresses from a three-month sample of emails from Hotmail, their system, i.e., AutoRE. However, there is a fatal weakness in that the three criteria, i.e., content, URL and domain name, are easily influenced by changes in spam messages and trends. In fact, spammers periodically change the contents of emails and domain names[12], and the active period of URLs is extremely short; 1 day: 45%, 2 days: 20%, 3 days: 7%[1]. Also, in our experiments, we confirmed that most URLs used in spam emails are unique. As a result, it could be said that these three criteria are not suitable for maintaining and improving the quality of clusters.

In [9][10], we carried out an experiment to examine the clustering of spam emails based on IP addresses resolved from URLs. In other words, we regarded two emails as the same cluster, called IP cluster, if their IP address sets resolved from URLs are completely identical. Although we demonstrated that its performance is better than that of the domain name and URL based clustering methods, there is a limitation in that the IP clusters contain lots of unrelated emails which were sent from different controlling entities, and whose URLs are connected to different types of Web pages. This is because there are many Web servers which are hosting a lot of Web sites on the same IP address, or many Web servers compromised by spammers or attackers are serving their Web sites. In this paper, we focus on dividing them belonging to the same IP cluster into the individual clusters.

3 Proposed method

3.1 Overall procedure

Figure 1 shows the overall procedure for the O-means clustering method proposed in this paper, and it is composed of the following 6 main parts.

- Header based clusterer: constructs HD clusters, denoted by HD_C_1, HD_C_2, …, HD_C_h, from given spam emails based on similarities between the email herder as described in Section **3.2** (①), and inserts them into O-means clusterer (②).

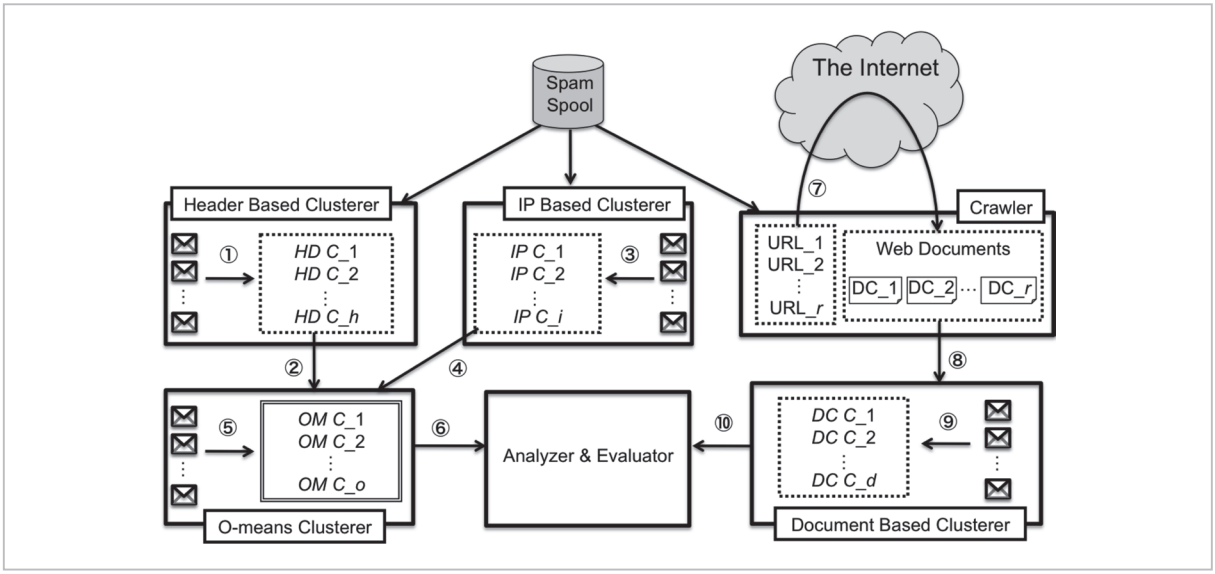


Fig.1 *Overall procedure of the O-means clustering method*

- IP based clusterer: resolves IP addresses from URLs within spam emails, generates IP clusters, denoted by IP_C_1, IP_C_2, …, IP_C_i, from spam emails using the resolved IP addresses as described in Section **3.3** (③), and inserts them into the O-means clusterer (④).
- O-means clusterer: generates OM clusters, denoted by OM_C_1, OM_C_2, …, OM_C_o, from spam emails based on the K-means clustering method as described in Section **3.4** (⑤), and inserts them into analyzer & evaluator (⑥).
- Crawler: accesses Web sites linked to URLs within spam emails, downloads their HTML content (⑦), and inserts them into document based clusterer (⑧).
- Document based clusterer: generates document clusters, denoted by DC_C_1, DC_C_2, …, DC_C_d, according to similarities found in Web document as described in Section **3.5** (⑨).
- Analyzer & evaluator: estimates the performance of the O-means clustering method as well as the 12 statistical features proposed in this paper, and analyzes spam sending systems and the URLs' destinations, i.e., Web servers, using the results from the O-means clusterer and the document based clusterer (⑥, ⑩).

3.2 Header based clusterer

In order to identify the relationship between spam sending systems, we leverage the email headers such as "From," "To" which indicate who the sender is and the receiver is, respectively, because the characteristics of the email headers depend on email client programs used for sending emails. In other words, since there are a lot of types of email headers of which some headers are essential, and others are optional, it can be said that if the email headers of two emails are not the same, then they are sent by different email client programs. In addition, in many cases, spammer and attackers use their individual spam sending programs, we can distinguish them by revealing their characteristics found in the email headers.

In our header based clusterer, we focus on three criteria, i.e., the types of email headers, their number and order, in order to classify spam emails into the HD clusters in which spam emails with the same email headers become members of the same HD cluster. Figure 2 shows its clustering process. During the clustering process, it first picks the email headers from each email (①) and extracts their types excluding overlapping (②). After that, it regards two emails as the same HD cluster if their number and order are completely identical (③).

3.3 IP based clusterer

Figure 3 shows the process of IP based clustering proposed in [9][10]. During the clustering process, it first picks unique URLs from the original URLs obtained from each email (①, ②). After that, it resolves IP address(es) from the unique URLs(③), and regards two emails as the same IP cluster if the IP address set resolved from the unique URLs is completely identical(④).

3.4 O-means clusterer

Figure 4 shows the clustering process of the O-means clustering method proposed in this paper. During the clustering process, we first extract the 12 statistical features from

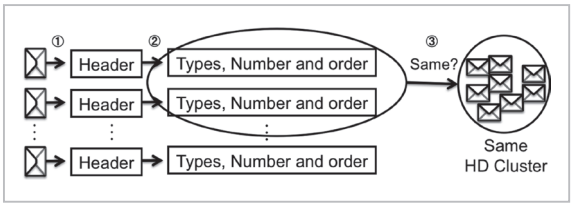


Fig.2 *Clustering process of the header based clusterer*

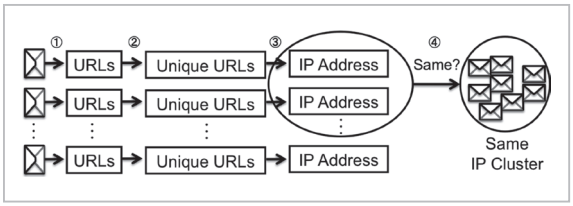


Fig.3 *Process of IP based clustering*

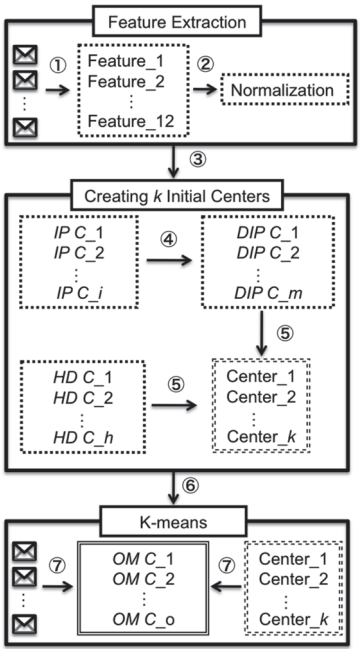


Fig.4 *Clustering process of the O-means clustering method*

each email as described in Section ***3.4.1*** (①) and normalize their values according to the normalization method described in Section ***3.4.2*** (②). We then create k initial centers using the IP clusters and the HD clusters as described in Section ***3.4.3*** (③, ④, ⑤, ⑥). Finally, we construct the OM clusters, i.e., OM_C_1, OM_C_2, …, OM_C_o, from spam emails using the k initial centers and the K-means clustering method (⑥, ⑦).

3.4.1 Feature extraction

Our feature extraction is based on the assumption that the spammers make spam emails, especially URLs, under a certain rule or pattern embedded in their email sending programs, even though the contents of emails, URLs and domain names are frequently changed[1][3][9][10][12]. Thus, there is a possibility that we are able to distinguish each spammer by characterizing his/her rule from spam emails and URLs. To this end, we define the 12 statistical features as shown in Table 1 and describe them using the example of an email as shown in Fig. 5.

In Figure 5, the email contains 7 lines representing the email headers and only one URL

Table 1 *Description of the 12 statistical features*

No.	Feature Name	Value
1	Size of emails	310
2	Number of lines	8
3	Number of unique URLs	1
4	Average length of unique URLs	57
5	Average length of domain names	17
6	Average length of path	10
7	Average length of query	22
8	Average number of key-value pairs	2
9	Average length of keys	5.5
10	Average length of values	4
11	Average number of dots(.) in domain names	1
12	Number of global top 100 URLs	0

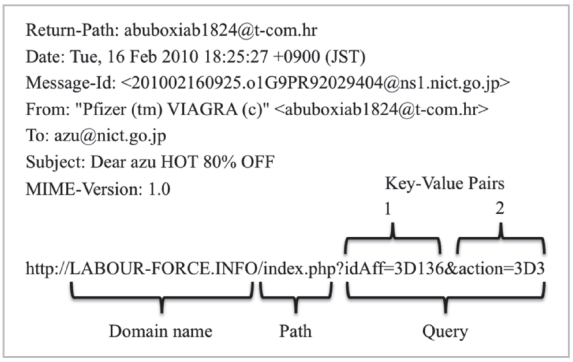


Fig.5 *Example of an email*

in its body part. From the email, we first compute the size of emails (i.e., 310) in bytes and the number of lines (i.e., 8). After that, we pick the unique URL from the email and divide it into 3 parts: domain name, path and query. Also, since the query part can contain multiple key-value pairs, we partition it into each key-value pair. By using those parts, we compute the values of 9 features (i.e., No. 3–No. 11) associated to the unique URL as shown in Table 1. Finally, we count the number of URLs that are most popular global top 100 Web sites provided from Alexa.com. The reason why we use this feature is that spammers may try to evade or confuse the URL based spam detection mechanism by crafting spam emails which contain legitimate URLs, especially the popular URLs such as "Google.com," "Yahoo.com," etc[1]. Therefore, we may be able to reveal the

characteristic of spammers by inspecting the using frequency of popular URLs.

3.4.2 Normalization

After we extracted the 12 statistical features from spam emails, we need to normalize their values, because each feature has a different scale. Our normalization method is basically based on [13]. Given a set of data instances (i.e., spam emails) which have the above 12 statistical values, we calculate the normalized values of each instance as follows.

$$normalized\_instnace[r] = \frac{original\_instance[r] - average[r]}{standard\_deviation[r]}$$

$$\textbf{subject to :}\ \forall r\ (1 \le r \le 12)$$

where $[r]$ is the rth feature, $average[r]$ and $standard\_deviation[r]$ are the average and the standard deviation of the 12 features obtained from all data instances, respectively. This normalization method means that for every feature value of each data instance, how far it is away from the average of the corresponding feature with respect to its standard deviation.

3.4.3 Creating k initial centers

(1) Motivation and strategy

After we obtained the above 12 normalized values from spam emails, we create k initial centers to be used in the next clustering phase, i.e., the K-means clustering method. In order to get the best performance out of the K-means clustering method, we need to choose the k initial centers as representatives of the actual k clusters within given spam emails. In the case of spam emails, each cluster can be defined as a group of spam emails which are sent from the same controlling entity, e.g., a botnet, and whose URLs are connected to the same Web page. In other words, if two emails are sent from different botnets, they should be members of different clusters, even if they share the same URL linked to the same Web page.

In order to reflect those clusters to the k initial centers, we leverage the IP clusters and the HD clusters. Our strategy is that we first merge the IP clusters into DIP (Duplicate IP) clusters, denoted by DIP_C_1, DIP_C_2, …, DIP_C_m, whose members (i.e., spam emails) share at least one IP address resolved from their unique URLs. This means that the members of each DIP cluster have a high possibility of having a close relationship to each other from viewpoint of URL destinations. However, the IP clusters contain lots of unrelated emails which were sent from different controlling entities, whose URLs are connected to different types of Web pages, even if they are hosting on the same Web server with the same IP address. Thus, we divide the DIP clusters into k groups based on the HD clusters whose members (i.e., spam emails) were sent from the same type of spam sending systems in terms of the types of the email headers, their number and order. As a result, it can be expected that if distinct spammers sent spam emails linked to the same Web page, then our method is able to discover each of them as a group. In addition, it is obvious that if a spammer is related to several different types of Web pages which are being hosted on different Web servers (i.e., the IP clusters) they can also be divided as different groups in our method.

(2) Making DIP clusters

Figure 6 shows the merging process of making the DIP clusters, and it is carried out by the following 7 steps.

① : feed all the IP clusters into a set of inter-

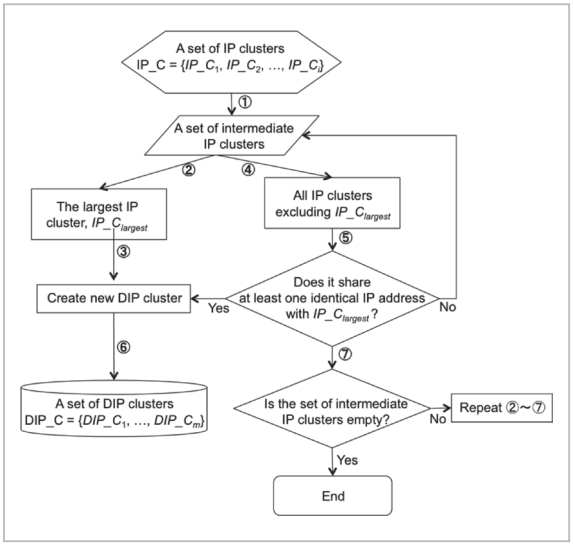


Fig.6 *Merging process of making the DIP clusters*

mediate IP clusters.

② : select the largest IP cluster, $IP\_C_{largest}$, from the set of intermediate IP clusters.

③ : create a new DIP cluster which contains only $IP\_C_{largest}$ as its initial member.

④ : select all IP clusters excluding $IP\_C_{largest}$ from the set of intermediate IP clusters.

⑤ : classify all IP clusters into two parts: if an IP cluster shares at least one identical IP address with $IP\_C_{largest}$, it becomes a member of the new DIP cluster, otherwise it returns to the set of intermediate IP clusters.

⑥ : add the new DIP cluster to the set of DIP clusters.

⑦ : repeat ②–⑦ unless the set of intermediate IP clusters is empty. Otherwise the merging process is terminated.

As a result, we can obtain the set of the DIP clusters, DIP_C_1, DIP_C_2, …, DIP_C_m.

(3) Creating k groups

Figure 7 shows the process of creating the k initial centers, and it is composed of the following 5 steps.

① : feed all the DIP clusters into a set of intermediate DIP clusters.

② : select a DIP cluster from the set of intermediate DIP clusters.

③ : partition its members, i.e., spam emails, into groups in which spam emails within the same group belong to the same HD cluster.

④ : add the new groups to the set of k groups.

⑤ : return to the step ② if the set of intermediate DIP clusters is not empty, otherwise the creation process is terminated.

From the above creation process, k groups, i.e., g_1, g_2, …, g_k are created. For the k groups, it computes the mean of their members, and regards them as the k initial centers.

After we create the k initial centers, we apply the K-means clustering method to spam emails. As the clustering result of the K-means clustering method, we are able to obtain the OM clusters, OM_C_1, OM_C_2, ⋯, OM_C_o. Note that we computed the distance between two emails using the Euclidean distance in our experiment. Given a pair of objects, e.g., $\mathbf{a}=\{a_1, a_2, \ldots, a_d\}$ and $\mathbf{b}=\{b_1, b_2, \cdots, b_d\}$, which are vectors in real d-dimensional space, $\Re^d$, then the Euclidean distance between $\mathbf{a}$ and $\mathbf{b}$, $d(\mathbf{a}, \mathbf{b})$, is as follows.

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{d}(a_i - b_i)^2}$$

3.5 Document based clusterer

In order to evaluate our clustering method, we clustered spam emails into the document clusters, DC_C_1, DC_C_2, …, DC_C_d, according to similarities found in Web pages linked to URLs. In order to examine the similarity between Web pages, we applied text shingling[1][10][14] to the corresponding Web pages. Figure 8 shows an example of text shingling where two Web documents, e.g., A and B, contain 50 (A_1, A_2, …, A_{50}) and 40 (B_1, B_2, …, B_{40}) words, the size of a shingle is 5. Since the size of a shingle is 5, we construct 46 shingles (e.g., $A_1A_2A_3A_4A_5$, $A_2A_3A_4A_5A_6$) and 36 shingles (e.g., $B_1B_2B_3B_4B_5$, $B_2B_3B_4B_5B_6$) from

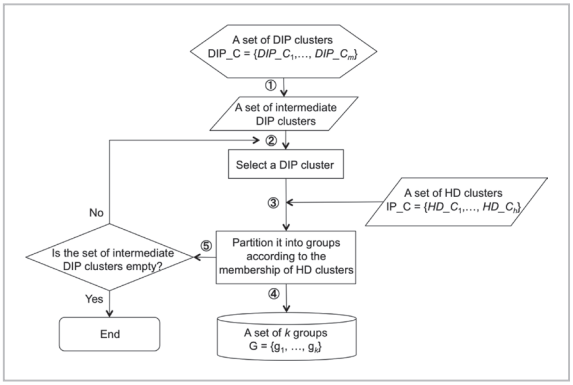


Fig.7 Creation process of the k initial centers

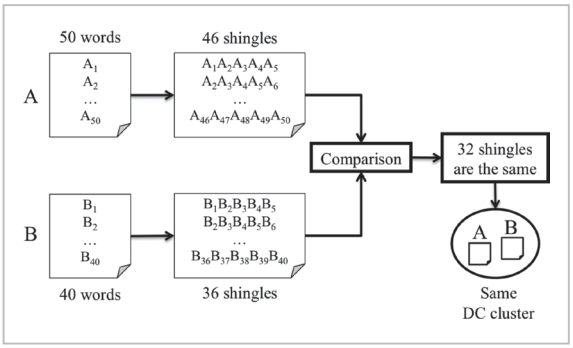


Fig.8 Example of text shingling

A and B, respectively, then similarity between A and B can be calculated as follows.

$$\frac{\textit{the number of the same shingles}}{\textit{the number of unique shingles in both A and B}}$$

In this example, if A and B share 32 shingles with each other and the threshold to determine whether they are the same cluster or not is 50%, then we regard them as members of the same DC cluster, because similarity between A and B is $32/(14+32+4)=64\%$ which is larger than 50%.

4 Experimental results

4.1 Double bounce emails

In double bounce emails, they have no valid email addresses associated with spam senders and receivers in their header. Figure 9 shows the overall process of double bounce emails. Assuming that there are three valid users whose email addresses are *a@user.com*, *b@user.com* and *c@user.com*. If a spammer sends a double bounce email in that its return-path address (i.e., forged@address.com) and recipient address (i.e., unknown@address.com) do not exist, to the target SMTP server (①), then the "user unknown" error message is exchanged between two SMPT servers (②, ③). From this situation, it could be said that the spammer intentionally forged his/her return-path address to conceal his/her activities and randomly generated a recipient address which does not exist in the real world. In the normal case, however, an email has at least one valid return-path address in its email header, even if a sender mistyped the recipient address to his/her email. In this context, double bounce emails can be regarded as pure spam, and therefore we use double bounce emails for our analysis data, i.e., spam emails.

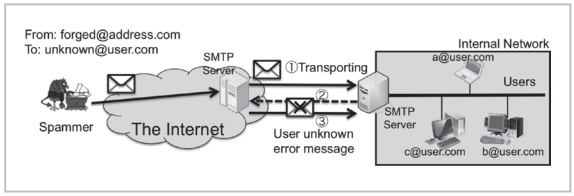


Fig.9 *Overall process of double bounce emails*

4.2 Description of experimental data

We collected 596,526 double bounce emails that arrived at our SMTP server for three weeks (Jan. 25th – Feb. 20th, 2010). Figure 10 shows their overall properties. Among all emails, we observed that 526,544 emails contained one or more URLs in their body, and the total number of URLs were 1,405,950 of which downloadable URLs and unique URLs numbered 1,296,120 and 1,048,158, respectively. Also, we found that 275,117 unique IP addresses were used for sending all of spam — this means each unique IP address was associated with only about 1.9 emails on average —, while the total number of unique IP addresses connected to URL destinations was only 1,643.

Figure 11 shows the national distribution of spam sending systems and URL destinations. From Figure 11, we can see that 10% and 8% of spam emails was sent from America and Brazil, respectively, but in total, our data found that spam was sent from a total of 204 countries. On the other hand, in the case of URL destinations (i.e., Web servers), 56% of them were located in China, but an additional 35% of Web servers were located in Korea (19%) and America (16%). These results show that the geographical distribution of spam send-

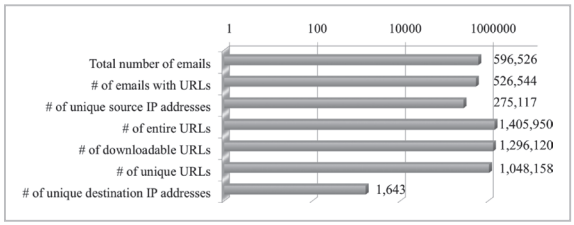


Fig.10 *Overall properties of experimental data*

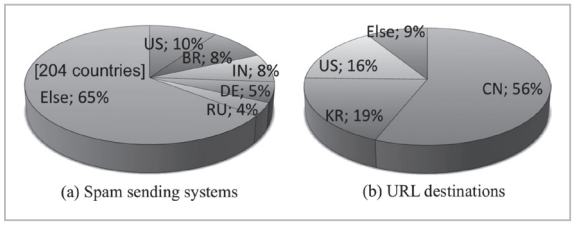


Fig.11 *National distribution of spam sending systems and URL destinations*

ing systems andWeb servers are quite different from each other. In other words, spam sending systems are widely distributed all around the world, but Web servers are concentrated in only three countries.

4.3 Results of document based clustering

In order to evaluate the performance of the O-means clustering method, we clustered double bounce emails into the DC clusters as described in Section **3.5**. During the clustering process, we set the size of a shingle to 5 and the threshold to 50%. This means that a shingle consists of 5 adjacent words, and if similarity between Web documents is larger than 50%, then they become members of the same DC cluster. In our investigation, we observed that there were 1,739 distinct Web documents whose hash values were different from each other among all Web documents crawled from 1,296,120 URLs.

We first grouped 1,739 Web documents under the above two conditions and found that there were 772 groups, i.e., there were 772 Web documents which contain distinct content from each other. Also, we observed that among 1,739 Web documents, 656 Web documents had no similar Web documents to the others, while 1,083 Web documents did. In our further investigation, we discovered that among the 772 groups, the most largest group has 156 similar Web documents as its members. Figure 12 shows similarity distribution of the 1,739 Web documents when they are assigned to one of the 772 groups. From Figure 12, it can be easily seen that similarities of most Web documents are close to either 0 or 100. This means that the threshold value (i.e., 50%) does not affect the performance of the document based clusterer.

We then grouped double bounce emails according to the membership with the 772 Web documents. In other words, if two double bounce emails share at least one URL — in many cases, it is not the same — which is linked to one of the 772 Web documents, then they become members of the same DC cluster. As a result, from experimental data, we obtained 772 DC clusters in which double bounce emails within the same DC cluster are similar to each other in terms of Web pages connected to URLs they contain.

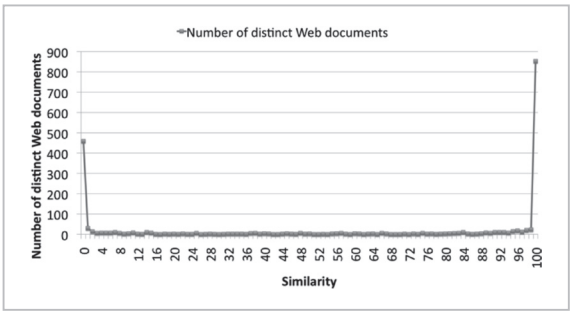


Fig.12 *Similarity distribution of 1,739 web documents*

4.4 Clustering results of the O-means clustering method

With respect to 526,544 emails with URLs, we constructed the HD clusters and the IP clusters as described in Sections **3.2** and **3.3**, respectively. The number of the HD clusters and the IP clusters were 1,062 and 1,204, respectively. Also, we merged the 1,204 IP clusters into the DIP clusters as described in Section ***3.4.3***, and obtained 900 DIP clusters. Using the 900 DIP clusters and the 1,062 HD clusters, we created 3,528 groups based on the creating process described in Section ***3.4.3***, and consequently computed 3,528 initial centers with the mean of their members. We then fed the 3,528 initial centers into the K-means clustering method which created the OM clusters, and consequently obtained 2,049 OM clusters.

In order to estimate the accuracy of the 2,049 OM clusters, we measured their accuracy using the DC clusters obtained from Section **4.3**. In this estimation, we first drew up a list of the DC clusters that contained at least one email from an OM cluster, and then we selected the certain DC cluster that shared the largest number of emails with the OM cluster. Finally, we calculated the accuracy of the OM cluster as follows.

$$\frac{\textit{the largest number of emails}}{\textit{the total number of emails in the OM cluster}}$$

From our evaluation, we observed that the average accuracy of the 2,049 OM clusters is 86.63% which is about 6% higher than that (i.e., 80.98%) of the IP based clusterer[10]. Figure 13 shows the accuracy of the IP based clusterer and the O-means clustering method with respect to the top 100 clusters — which are responsible about 98% and 86% of double bounce emails used in our experiment, respectively — in terms of their size, i.e., the number of members. Specifically, we can observe that the O-means clustering method is superior to the IP based clusterer in the 10 clusters among the top 12 clusters. The accuracy of the top 10 OM clusters is shown in Table 2.

Table 2 shows statistical information of spam sending systems and URL destinations for the top 10 OM clusters. From Table 2, we can see that each OM cluster has a lot of distinct spam sending systems, i.e., 4,730–27,476, which represent the size of botnets or systems connected closely with each other and that they are distributed in many different countries, i.e., 90–151. While in the case of URL destinations, we can see that the number of unique IP addresses is only 2–5 and that they are located in 1–3 countries. In fact, in our investigation, we observed that they are distributed in one of three countries, i.e., China, Korea and America. In Table 2, the most important thing to note is that each OM cluster, especially the top 8 OM clusters, has many distinct URLs and domain names. This means that spammers

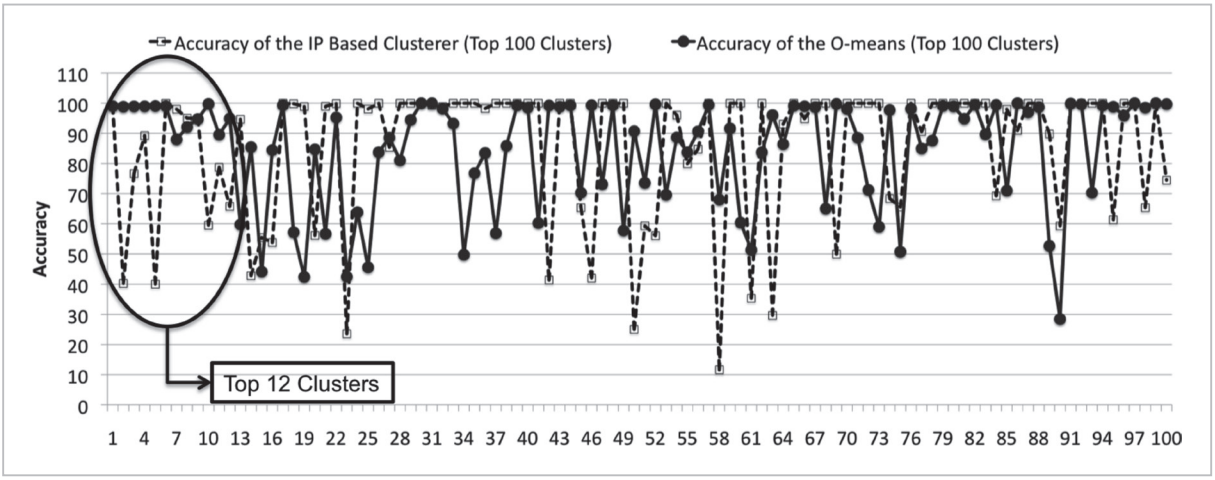


Fig.13 *Performance comparison between the IP based clustering method and the O-means clustering method with respect to top 100 clusters*

Table 2 *Statistics of the top 10 OM clusters*

	OM Cluster ID (Top 10 Clusters)									
	1079	82	24	12	19	22	16	10	129	1193
Accuracy	99.03%	98.84%	98.95%	98.98%	99.08%	99.05%	87.99%	92.12%	94.74%	99.76%
# of emails	31,790	22,857	20,230	17,817	9,850	9,409	9,369	9,013	8,998	6,772
# of unique source IP addresses	27,476	20,176	18,244	16,260	9,229	8,901	7,681	7,368	5,731	4,730
# of unique source countries	108	108	106	106	92	90	94	91	149	151
# of entire URLs	**95,370**	**68,619**	**60,688**	**53,447**	**29,548**	**28,225**	**56,214**	**54,073**	8,998	6,772
# of unique URLs	**94,805**	**68,198**	**60,294**	**53,151**	**29,532**	**28,068**	**56,008**	**53,958**	12	5
# of unique domain names	**94,805**	**68,198**	**60,294**	**53,151**	**29,532**	**28,068**	**56,008**	**53,958**	12	5
# of unique destination IP addresses	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**5**	5	2
# of unique destination countries	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	3	2

or attackers frequently change their URLs and domain names. Therefore, in the case of the existing domain name and URL based clustering methods, it could be said that the clustering accuracy is extremely low, because most URLs are unique as well as they have unique domain names.

4.5 Feature selection method

In this section, we present our feature selection method based on heuristic analysis of spam emails. In our method, we first selected the top 10 clusters from the original 772 clusters obtained from the document based clustering described in Section **3.5**. Table 3 shows the the statistical information of the top 10 clusters (see Section ***4.5.1***). Using the top 10 clusters, we evaluated 12 features heuristically and identified their degrees of contribution for improving the clustering accuracy of spam emails (see Section ***4.5.2***).

4.5.1 Top 10 clusters

Using the document based clustering described in Section **3.5**, we obtained 772 clusters and observed that there are 10 large clusters which are responsible for about 90% of spam emails with URLs arriving at our SMTP server. Table 3 shows their statistical information. From Table 3, we can see that each cluster has a lot of distinct spam sending systems, i.e., 3,061–169,149, which represent the size of botnets or systems connected closely with each other and that they are distributed in many different countries, i.e., 101–192. While in the case of URL destinations, we can see that the number of unique IP addresses is only 3–164 and that they are located in 2–29 countries. Also, we can observe that there are two patterns of URLs and domain names: in the case of clusters 1, 3 and 5, spammers or attackers frequently change their URLs and domain names, while in the case of the others, they almost always used the same URLs and domain names.

4.5.2 Selecting significant features

Although we defined the 12 statistical features to calculate similarity between spam emails and showed that the clustering accuracy is superior to the existing research, we need to consider that the relationships among the 12 statistical features are not independent from each other, and thereby there may exist several redundant features which do not contribute to the improving of the clustering accuracy. Furthermore, considering we have to deal with a large amount of spam emails and analyze them effectively, it is needed to extract significant features from the original 12 features, so that we are able to reduce the analysis time of spam emails while maintaining the high clustering accuracy.

In order to discover a set of optimized features, it is best to investigate all the combinations of the 12 features: namely, the number of

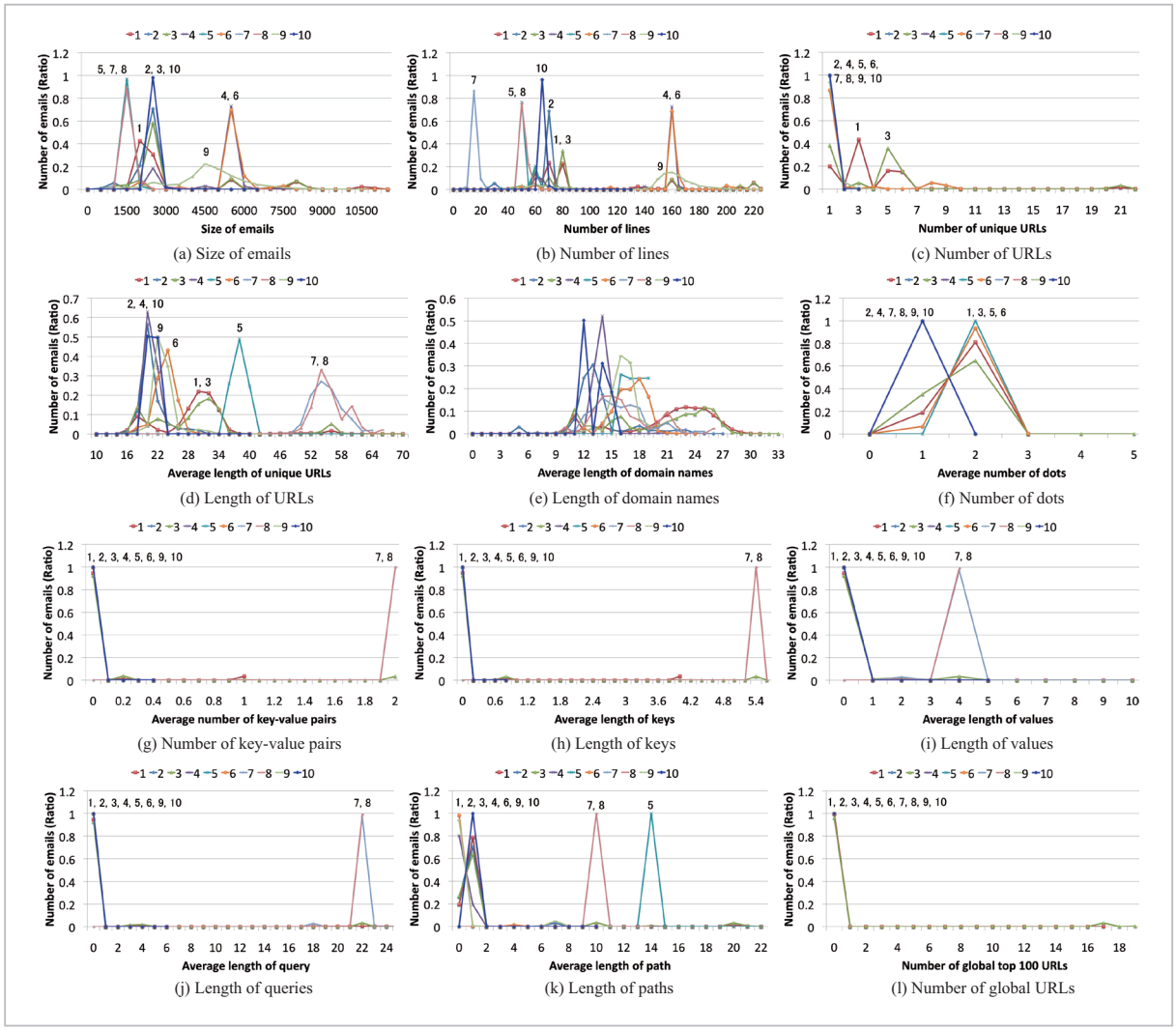
Table 3 *Statistics of top 10 clusters*

ID	Top 10 Clusters									
	1	2	3	4	5	6	7	8	9	10
A	286,024	30,205	22,696	21,061	17,393	10,741	9,807	7,649	5,601	5,265
B	169,149	19,102	16,881	10,480	4,003	4,828	6,101	5,241	4,958	3,061
C	192	177	163	162	101	120	120	127	133	137
D	951,565	30,209	77,794	21,061	17,397	18,363	9,806	7,649	5,919	5,254
E	891,795	177	68,584	94	16,790	6,306	102	67	155	8
F	890,022	157	68,344	79	16,757	6,300	97	63	135	6
G	80	11	164	14	100	21	3	3	13	3
H	11	6	29	5	30	4	2	3	4	3

A: Num. of emails, B: Num. of unique source IP addresses, C: Num. of unique source countries, D: Num. of entire URLs, E: Num. of unique URLs, F: Num. of unique domain names, G: Num. of unique destination IP addresses, H: Num. of unique destination countries

all cases to be estimated is $\sum_{i=1}^{12} {}_{12}C_i$ but doing so is extremely time-consuming. Thus, we carried out an alternative experiment, which is based on the following heuristic analysis of the top 10 clusters.

1. In each cluster, the number of spam emails that have the same value is counted with respect to all of the 12 features.
2. The number of spam emails is scaled to $[0, 1]$ according to the size of each cluster, because the size of the top 10 clusters is different from each other.
3. Made 2-dimensional graphs for each feature where the horizontal axis indicates the values of each feature and the vertical axis indicates the ratio of the number of spam emails within each cluster as shown in Fig. 14.
4. Using the results in Fig. 14, we first selected three significant features: "Size of emails" (14(a)), "Number of lines" (14(b)), "Length of URLs" (14(d)), because most clusters can be distinguished from each other by using their values. In other words, since spam emails in each cluster have different value distribution in these three features, if we use the three features to calculate similarity between spam emails it is possible to accurately partition them into clusters where members within the same cluster have similar values with each other.
5. Second, we excluded "Number of global URLs" (14(l)) from the list of significant features, because the top 10 clusters have



(a) Size of emails (b) Number of lines (c) Number of URLs
(d) Length of URLs (e) Length of domain names (f) Number of dots
(g) Number of key-value pairs (h) Length of keys (i) Length of values
(j) Length of queries (k) Length of paths (l) Number of global URLs

Fig.14 *Value distribution of 12 statistical features with respect to top 10 clusters*

similar values in this feature.

6. Third, we did not regard "Length of domain names" (14(e)) as a significant feature, because it is unable to help distinguish each cluster at all: there is no cluster whose value distribution is independent from others.
7. Fourth, we also excluded 6 features, i.e., "Number of URLs" (14(c)), "Number of key-value pairs" (14(g)), "Length of keys" (14(h)), "Length of values" (14(i)), "Length of queries" (14(j)), "Length of paths" (14(k)) from the list of significant features, because they helped distinguish clusters 7, 8, 1, 3 and 5 from the others, but it is obvious that those clusters can also be obtained by using the three significant features in step 4.
8. Finally, since it is impossible to distinguish cluster 4 from cluster 5 using the three significant features in step 4, we added "Number of dots" (14(f)) to the list of significant features as it was able to separate them.

As a result, we selected four significant features, i.e., "Size of emails," "Number of lines," "Length of URLs", and "Number of dots". Note that our method is a supervised feature selection, since it uses the label information of top 10 clusters.

4.5.3 Performance evaluation

In order to demonstrate the effectiveness of the four significant features, we examined their clustering accuracy and time complexity. As a clustering algorithm, we used an optimized spam clustering method, called O-means based on the K-means clustering method, which is one of the most widely used clustering methods. The evaluation results are shown in Table 4. From Table 4, we can see that almost the same clustering accuracy (i.e., 86.33%) was yielded by using only these four significant features. In addition, we can see that execution time was drastically reduced as a result of using only these four significant features, enabling us to analyze spam emails more effectively. This measurement was performed on a machine running an Intel Core 2 Duo 2.8GHz CPU with 3 GB of RAM, and our program was written in the Perl programming language and Mysql.

Table 4 *Comparison of the clustering accuracy and time complexity*

	4 significant features	All features
Clustering Accuracy	86.33%	86.63%
Execution Time (sec.)	6,124	28,772

5 Discussion

In order to evaluate the performance of the O-means clustering method, we constructed the document clusters (i.e., DC_C_1, DC_C_2, …, DC_C_d) using text shingle technique as shown in Section **3.5**. Although there are a lot of techniques (e.g., Levenshtein distance) for estimating similarities in text documents, in our method we used the text shingling technique to compare the similarity of Web pages downloaded from URLs, because it was developed to measure and compare the similarity of Web pages, and its effectiveness was verified in many approaches[1][3][10][14][16][17]. Furthermore, our experimental results shown in Fig. 12 also demonstrate that 1,739 Web documents were well classified according to similarities among them by using this text shingling technique.

In our experiments, we evaluated the O-means clustering method using 3 weeks of double bounce emails that arrived at our SMTP server. In order to evaluate the Omeans clustering method more accurately, it is better to analyze a longer period of double bounce emails and Web pages linked to URLs. However, there is a practical problem which makes this difficult: Web crawling is an intrusive process that might let spammers believe certain groups of users are more vulnerable to spam emails and thus send more spam to them in the future[1]; the time complexity for computing similarities among all Web documents downloaded from URLs increases exponentially.

Similar to our experiments, therefore, most of existing approaches evaluated their systems using a short period of spam emails. In fact, in [1][7][8], they used only one month, one week and 9 days of spam emails, respectively. Nevertheless, they succeeded in identifying of a number of botnets and spam clusters. As one of the reasons for such a success, it could be considered that the active time of recent bots and the effective lifetime of a single spam message is extremely short[2][3]. In this context, it could be concluded that our experiments were carried out in a reasonable manner.

6 Conclusion

In this paper, we have proposed an optimized spam clustering method, the O-means clustering method, based on the K-means clustering method[11], which is one of the most widely used clustering methods. The O-means clustering method improves its performance by overcoming the shortcomings of the K-means clustering method: its clustering result depends on the chosen k initial centers, and it is very difficult to predefine the proper number of clusters, i.e., k. By examining three weeks of spam gathered in our SMTP server, we observed that the accuracy of the O-means clustering method is about 87% which is superior to the previous clustering methods. In addition, we have defined new 12 statistical features to compare similarities between spam emails, and we have proposed a feature selection method to identify a set of optimized features which makes the Omeans clustering method more effective. With our method, we identified 4 significant features which yielded a clustering accuracy of 86.33% with low time complexity.

Although we only focused on the improvement of the clustering accuracy in this paper, the spam clusters obtained from the O-means clustering method can be utilized for analyzing spam based attacks more accurately. Therefore, we need to carry out more practical analysis for spam based attacks in our future work, so that we are able to identify the infrastructure of spam sending systems and malicious Web servers, and how they are grouped and correlate with each other. Further, it is worth to investigate how much the spam clusters can be contributed to the time reduction of analyzing Web pages lined to URLs. We also need to consider the merging of the OM clusters generated by the O-means clustering method, because in the creating process of the k initial centers, we regarded two spam emails as different clusters, if they were sent from different controlling entities, even though their URLs are linked to the same Web page. By merging the OM clusters according to their similarity, we are able to identify the relationship between the OM clusters; it could be expected that we are able to identify a group of the OM clusters whose emails share URL(s) linked to the same Web page. It is also a challenging task to devise additional statistical features and to evaluate them with respect to all their combinations, so that we are able to obtain an optimized set of statistical features for the clustering of spam emails. Finally, we will evaluate the O-means clustering method with a larger data set of spam obtained from various domain sources in order to analyze spam based attacks more effectively.

References

1 Xie, Y., Yu, F., Achan, K., Panigrahy, R., Hulten, G., and Osipkov, I., "Spamming botnets: signatures and characteristics," ACM SIGCOMM Computer Communication Review, Vol. 38, No. 4, October 2008.

2 Ramachandran, A. and Feamster, N., "Understanding the network-level behavior of spammers," Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications, Vol. 36, No. 4, September 11–15, Pisa, Italy, 2006.

3 Anderson, D. S., Fleizach, C., Savage, S., Voelker, and G. M., “Spamscatter: Characterizing Internet Scam Hosting Infrastructure,” Proceedings of the USENIX Security Symposium, Boston, 2007.
4 Jennings, R., “The global economic impact of spam, 2005 report,” Technical report, Ferris Research, 2005.
5 Kawakoya, Y., Akiyama, M., Aoki, K., Itoh M., and Takakura, H., “Investigation of Spam Mail DrivenWeb-Based Passive Attack,” IEICE Technical Report, ICSS2009-5(2009-5), pp. 21–26, 2009.
6 Jungsuk Song, Daisuke Inoue, Masashi Eto, Suzuki Mio, Satoshi Hayashi, and Koji Nakao, “A Methodology for Analyzing Overall Flow of Spam-based Attacks,” 16th International Conference on Neural Information Processing(ICONIP 2009), LNCS 5864, pp. 556–564, Bangkok, Thailand, 1–5 December 2009.
7 Li, F. and Hsieh, M.H., “An Empirical Study of Clustering Behavior of Spammers and Group-based Anti-Spam Strategies,” Proceedings of 3rd Conference on Email and Anti-Spam (CEAS), pp. 21–28, Mountain View, CA, 2006.
8 Zhuang, L., Dunagan, J., Simon, D. R., Wang, H. J., and Tygar, J. D., “Characterizing botnets from email spam records,” Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, No. 2, pp. 1–9, San Francisco, 2008.
9 Jungsuk Song, Daisuke Inoue, Masashi Eto, Hyung Chan Kim, and Koji Nakao, “Preliminary Investigation for Analyzing Network Incidents Caused by Spam,” The Symposium on Cryptography and Information Security (SCIS 2010), Takamatsu, Japan, 19–22 January, 2010.
10 Jungsuk Song, Daisuke Inoue, Masashi Eto, Hyung Chan Kim, and Koji Nakao, “An Empirical Study of Spam : Analyzing Spam Sending Systems and Malicious Web Servers,” 10th Annual International Symposium on Applications and the Internet (SAINT 2010), Seoul, Korea, 19–23 July 2010.
11 MCQUEEN, J, “Some methods for classification and analysis of multivariate observations,” Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297, 1967.
12 John, J. P., Moshchuk, A., Gribble, S. D., and Krishnamurthy, A., “Studying spamming botnets using Botlab,” Proceedings of the 6th USENIX symposium on Networked systems design and implementation, pp. 291–306, Boston, April 22–24, 2009.
13 L. Portnoy, E. Eskin, and S. Stolfo, “Intrusion Detection with Unlabeled Data Using Clustering,” Proceedings of ACM CSS Workshop on Data Mining Applied to Security, 2001.
14 D. Fetterly, M. Manasse, M. Najork, and J. L.Wiener, “A large-scale study of the evolution of web pages,” Softw. Pract. Exper., 34(2), 2004.
15 Huan Liu and Lei Yu, “Toward integrating feature selection algorithms for classification and clustering,” IEEE Transactions on Knowledge and Data Engineering, Vol. 17, pp. 491–502, 2005.
16 A. Broder, “ On the Resemblance and Containment of Documents,” Proceedings of the Compression and Complexity of Sequences (SEQUENCES ’97), pp. 21–29, June 11–13, 1997.
17 N. SHIVAKUMAR and H. GARCIA-MOLINA, “Finding near-replicas of documents and servers on the web,” Proceedings of the First International Workshop on the Web and Databases (WebDB’98), pp. 204–212, March 1998.

(Accepted June 15, 2011)

SONG Jungsuk, *Ph.D.*

Researcher, Cybersecurity Laboratory, Network Security Research Institute

Network Security, Spam Analysis, IPv6 Security