Establishing Phishing Provenance Using Orthographic Features

Liping Ma, John Yearwood, Paul Watters
Internet Commercial Security Laboratory (ICSL)
Centre for Informatics and Applied Optimization
Graduate School of Information Technology and Mathematical Sciences
University of Ballarat, Australia
Email: (l.ma, j.yearwood, p.watters)@ballarat.edu.au

***Abstract*—After phishing message detection, determining the provenance of phishing messages and websites is the second step to tracing cybercriminals. In this paper, we present a novel method to cluster phishing emails automatically using orthographic features. In particular, we develop an algorithm to cluster documents and remove redundant features at the same time. After collecting all the possible features based on observation, we adapt the modified global k-mean method repeatedly, and generate the objective function values over a range of tolerance values across different subsets of features. Finally, we identify the appropriate clusters based on studying the distribution of the objective function values. Experimental evaluation of a large number of computations demonstrates that our clustering and feature selection techniques are highly effective and achieve reliable results.**

***Index Terms*—Clustering, feature selection, feature elimination, modified global k-means**

I. Introduction

Phishing is a criminal activity using social engineering techniques to fraudulently acquire personal information, such as credit card and bank details or passwords. In a typical phishing attack, the phishers send a large number of malicious emails that pretend to come from a legitimate organisation, typically a financial institution such as a bank or insurance company. The emails often urge the users to update their personal information in a fake or spoofed website which was created to be almost identical to the legitimate organization's website so that the phishers may steal the identities of the users via the website. Once this information is acquired, the phishers may use a person's details to create fake accounts in a victim's name, ruin a victim's credit, or even prevent victims from accessing their own accounts.

A study by academics at Harvard and Berkeley universities reported in The Register, reveal that 23% of users only look at the content of sites when deciding whether they are legitimate. A survey of Gartner [14] on phishing attacks shows that approximately 3.6 million users in the United States suffered losses caused by phishing, totalling approximately US$3.2 billion. Especially, the number of individual victims rose from 2.3 million in 2006 to 3.6 million in 2007, which is a 56.5% increase.

The damage caused by phishing ranges from loss of access to email to substantial financial loss. This style of identity theft is becoming more popular, because of the ease with which unsuspecting people often divulge personal information to phishers, including credit card numbers, passwords, and so on. There are also fears that identity thieves can obtain some such information simply by accessing public records. Phishing has become more and more complicated and sophisticated so that phishers can bypass the filter set by current anti-phishing techniques and cast their bait to customers and organizations.

According to the report of the Anti-Phishing Working Group [12] the longest time for a phishing site to exist is thirty-one days, and the average time online for a phishing site is shorter than five days. Phishers may use different URLs at different times. This introduces the issue of how to forensically identify the provenance of phishing emails, i.e., how to determine whether phishing emails and their associated websites come from the same group, if the URLs are not the same. One possible solution is to cluster or group the emails.

Clustering is typically implemented as fully automated, unsupervised learning algorithms and similar ones are grouped whilst different ones find different groups. The algorithm is usually given input documents and features selected [5], [17]. However, recently researchers have focused on allowing a user to provide limited information (i.e. text feature) to improve clustering quality. As in traditional information retrieval, these clustering approaches focus on text clustering, input information consists of text features. For example, users may supply a few keywords per cluster and a class hierarchy to generate preliminary labels to create an initial text classifier for the cluster [4], [15], [16]. Although additional input involves background knowledge to enrich the set of features that describe each document [11] such as ontology, orthographic features have rarely been considered (refer to Section III for a definition).

Most phishing emails are largely similar in wording, especially the most important terms, such as "security", "expire", "unauthorized", "account", "login" and so on. Such terms are useful to classify if an email is a phishing email, however, this content may not be so helpful in determining the provenance

978-1-4244-4626-1/09/$25.00 2009 Crown

of the email. This important next step in the forensic process requires an analysis of other features which may uniquely identify a particular group's style.

Since the semantic terms are largely similar in such documents, the orthographic features may differentiate between emails better. For example, phishers often intend to lure customers in different ways. They may redirect the customers to a website, require the customers to login using a form, create invisible links which the customers may click on accidently. There are many different such styles (orthographic features) in each document. Orthographic features are styles that are used to describe presentation of segments in documents including lexical, syntactic, structural and content features (details refer to Section III). This paper presents work on clustering phishing emails using orthographic features, with the hypothesis that such features are the most effective in identifying the provenance of such messages. The experiment in Section V shows that orthographic features are good discriminators in clustering the same type of documents.

The rest of the paper is structured as follows: Section II places our work in the context of existing work in anti-phishing and text clustering; Section III gives the details of feature collection; Section IV considers how to select features and identify clusters; Section V provides experimental results on the effectiveness of the clustering; Section VI is the conclusion which summarises the work and directions for future work.

II. Related work

Text clustering is useful, since it enables us to group documents automatically without any prior knowledge and supervision. There has been a lot of work in the information retrieval and machine learning communities on the problem of text clustering. Various methodologies have recently been developed for document classification and representation to assist in anti-phishing [10], [18], [19], [21], [25], [29], [22], [23], [20].

A. Anti-phishing methods

There are only a few research efforts that focus entirely on tackling the problem of phishing attacks. Also, most of the existing work focuses on the techniques to predict whether an email is malicious [7], [10], [18], [19], [21], [1]. Suspect emails are signed as "spam" email and the users are protected from accessing emails. This type of solution falls in the problem of text classification which classifies emails to two categories: normal emails or phishing emails.

AntiPhish [18] is a browser extension which is used to protect inexperienced users against spoofed web site-based phishing attacks. AntiPhish is a plug-in tool which keeps track of users' sensitive information and prevents this information from being passed to a web site that is considered as untrusted. A text classification algorithm is responsible for identifying whether a web site is a phishing site based on addresses used in a form. In detail, it compares a legitimate URL and IP address with URL the page actually locates. AntiPhish focuses more on tracking sensitive information provided by a user. In contrast, [29] identified a website as a suspect phishing site when the visual similarity value is above a pre-defined threshold.

Another widely-deployed technique is based on using a blacklist of phishing domains to force the browser to refuse to visit, such as PwdHash [6], [25] and SpoofGuard [25], [26] by Stanford University. However, it is currently unclear how effective such blacklisting approaches are in mitigating phishing attacks in reality, given the use of FastFlux and other technologies to rapidly change hosting locations.

There are many existing machine methods for text classification, such as the decision tree used by [21]. Text classification aims to automatically categorise text documents into predefined classes or types based on their contents [28]. However, when dealing with phishing there are a priori no identified classes. However, fewer researchers work on identified phishing emails.

B. Traditional clustering methods

The use of machine learning techniques to iterate feature selection has received more and more attention. Early approaches have been developed on supervised learning for either dimensionality reduction or increasing feature quality, such as feature selection in text classification. Since information available on the web is more and more complicated and has increased dramatically, the need for unsupervised learning has risen.

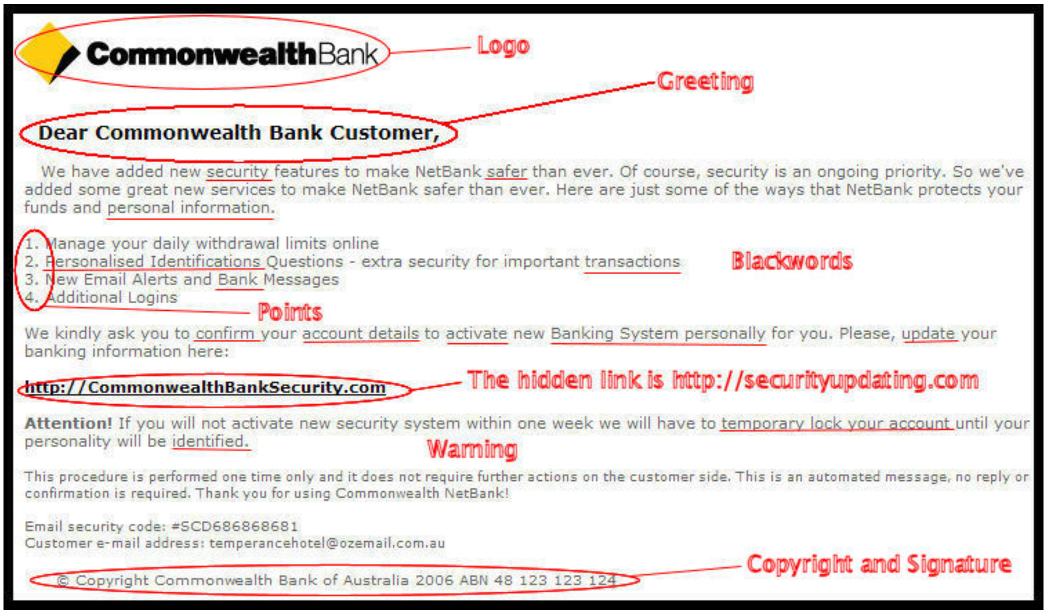
Roth, et. al [27] developed a system to implement a wrapper strategy of feature selection for text clustering. The features are directly selected by optimizing the discriminative power of the partitioning algorithm used. However, the experiment only resulted in clustering two clusters which is not suitable to implement multiple clusters. Also, it could be considered as a text classification problem which is supervised.

Dash and Liu [8], [9] selected features based on a ranking algorithm. All the features firstly are ranked based on the importance of clustering using entropy-based ranking measure. Then the features are added into the cluster one by one from the most important feature to the least important one. When the effectiveness of the cluster does not increase, the selection ends by the values of the objective function. An assumption here is that the ranking process must be correct, therefore, important features may be ignored when the ranking process is not reliable.

C. Problems and solutions in this paper

The problem and solution presented in this paper differ from existing systems as follows:

1) Application domain: Most existing systems work on identifying phishing emails, while we aim to cluster emails which were already identified as phishing emails by their origin.
2) Feature space: Phisher emails are largely similar in content. Therefore, we believe that the orthographic features may be more informative than the semantic



CommonwealthBank

Dear Commonwealth Bank Customer,

We have added new security features to make NetBank safer than ever. Of course, security is an ongoing priority. So we've added some great new services to make NetBank safer than ever. Here are just some of the ways that NetBank protects your funds and personal information.

1. Manage your daily withdrawal limits online
2. Personalised Identifications Questions - extra security for important transactions
3. New Email Alerts and Bank Messages
4. Additional Logins

We kindly ask you to confirm your account details to activate new Banking System personally for you. Please, update your banking information here:

http://CommonwealthBankSecurity.com

Attention! If you will not activate new security system within one week we will have to temporary lock your account until your personality will be identified.

This procedure is performed one time only and it does not require further actions on the customer side. This is an automated message, no reply or confirmation is required. Thank you for using Commonwealth NetBank!

Email security code: #SCD686868681
Customer e-mail address: temperancehotel@ozemail.com.au

© Copyright Commonwealth Bank of Australia 2006 ABN 48 123 123 124

Fig. 1. Commonwealth Phishing Emails 18 May 2006

features in such application. The work presented in this paper focuses on orthographic features.

3) Feature elimination and selection: The quality of features used in clustering determines greatly the effectiveness of a cluster. We design and implement a novel algorithm that adapts the modified global k-mean method and largely evaluates the objective function values across different subsets of features. This algorithm reinforces the clustering by eliminating “noisy” features so that only good features are selected.
4) Cluster effectiveness: Clustering is an unsupervised learning process so that it is not easy to judge whether a cluster is appropriate. We develop a method to identify clusters by analysing a large number of figures which indicate the relationship among objective function values, tolerance values and number of clusters based on features of various level and subset. From the analysing, we are able to identify confident clusters.

III. Document presentation

The standard document representation used in clustering is the vector space model [24]. In this model, each document is represented by a vector of (term, value) pairs observed in the document. Then a collection of a document can be represented as a set of document vectors or a matrix.

A. *Feature collection and definition*

A phishing email usually contains multimedia information, including image and text, where the text information may contain plain text, HTML, URLs, scripts, styles and so on. However, the information cannot be recognized by a system directly, rather it needs to be characterized according to the needs of the system.

As discussed in Section II, phishing emails are largely similar in content. Therefore, we believe that orthographic features are more informative than the semantic features in such an application. The orthographic features mainly consist of style characteristics that are used to convey the role of words, sentence or section and other useful description of the content. They are defined manually based on observation.

Since an email body is often loose in structure, parsing email content is more difficult than parsing the heading part of the email. The work presented in this paper focuses on the orthographic structure of the email content.

The basic orthographic features that are considered include HTML features, size of document and other elements. The original orthographic features collected in our system are described as following:

1) size of email: size of text body and html body of an email.
2) text content: whether an email has content [1].
3) vlinks: number of visible links in an email.
4) same_vHyLink: whether existing a visual link is directed to the same hyperlink in an email.
5) greetings: whether an email contains greeting line.
6) signature whether an email contains signature at the end.
7) html content: whether an email contains HTML content.
8) script: whether an email contains javescript and other scripts.
9) table: whether an email contains tables.
10) image: number of images in an email.
11) link: number of hyperlink in an email.
12) form: wether an email contains form.
13) fake tags: number of faketages in a email.

[1] Some phishing emails contain only image

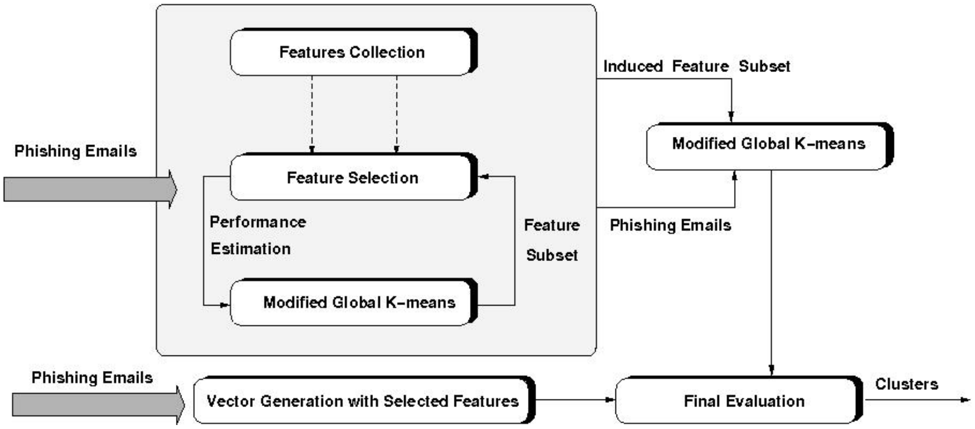


Fig. 2. Feature selection using modified global k-mean

By implementing the algorithms described in Figure 3 (Section IV), we identified thirteen useful orthographic features as shown in Table I. Most of the features are self-explanatory. The “script” feature indicates whether an email contains any script, while the “fake tags” indicate the number of fake tags in the email. The “greetings” indicates whether an email contains a greeting line such as “Dear customer” or “Dear User”. The “signature” indicates the content at the end of an email, including “Copyright” or “ABN Number” of a bank which to be used to make the email is more likely coming from a legitimate organization.

Figure 1 shows an example of phishing email delivered on 18th May 2006. The email contains a logo, a greeting line, text content, a hyperlink, signature, etc. We discovered that the hidden link of the visible hyperlink are different. We have also identified other useful features such as blackword list, dot points, paragraph, and warning, that are useful in phishing email prediction and phishing email clustering with structural features.

B. Document presentation and feature normalization

After features are defined, we developed a set of methods to extract all thirteen possible useful features from each email. Let $D = \{d_1, d_2, \ldots, d_{|D|}\}$ denote all the documents and $V = \{v_1, v_2, \ldots, v_{|V|}\}$ be the feature vector space. Where $|D|$ and $|V|$ are the number of document and size of feature vector respectively. Let a_{ij} be the value of jth feature of ith document. Therefore, the presentation of each document is $A_i = (a_{i1}, a_{i2}, \ldots, a_{i|V|})$, and each document is $A = \{a_{ij}\}$ where $i = 1, 2, \ldots, |V|; j = 1, 2, \ldots, |D|$.

The values of all features are numerical but in a different range. For example, the size of a document could be thousands byte while the number of images may be under five. To treat all the original features as equally important, the value of each feature needs to be normalized before the clustering process. Feature values are normalized using the quotient of the actual value over the maximum value among the feature so that numerical values are limited to the range $[0, 1]$.

Let b_{ij} be the value of jth normalized feature of ith document, $b_{ij} = \frac{a_{ij}}{\max\{a_{kj}, k = 1, 2, \ldots, |D|\}}$, the ith document $b_i = (b_{i1}, b_{i2}, \ldots, b_{i|V|})$, all the documents are represented as $B = \{b_{ij}\}$. The normalized feature values are shown in Table I.

Name of Feature	Normalized Values
size of email	[0..1]
text content	0 or 1
vlinks	[0..1]
same_vHyLink	0 or 1
greetings	0 or 1
signature	0 or 1
html content	0 or 1
script	0 or 1
table	[0..1]
image	[0..1]
links	[0..1]
form	0 or 1
fake tags	[0..1]

TABLE I
ORTHOGRAPHIC FEATURES

IV. Feature Selection

Feature collection gives us a set of possible orthographic features. However, not every feature is effective as a discriminator for the purposes of provenance determination. Therefore, it is necessary to select a relevant subset from the feature set upon which to focus our attention, while ignoring the rest. The problem of feature selection is that of finding a subset of the original features of a data set, such that an iteration algorithm which runs on data containing only these features generates a cluster with the highest possible accuracy. Meanwhile, all the clusters are also identified.

As described in Section III-A, the length of the orthographic representation is not large. Therefore, the purpose of feature selection in this work is to increase the quality of the feature vector for better discrimination and reduce noise rather than dimensionality reduction.

```
Input:            All possible features $V = \{v_1, v_2, \ldots, v_{|V|}\}$, phishing emails $D = \{d_1, d_2, \ldots, d_{|D|}\}$
Output:           Optimized set of features $\hat{V} = \{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_{\hat{V}}\}$
                  Number of clusters, Cluster $C$
begin
pre-precessing:   $M = \{m_1, m_2, \ldots, m_{|V|}\} = \{0, 0, \ldots, 0\}$
                  for all $D_i \in D$ do
                      for all $f_j \in F$ do $a_{ij} = f_j(Di)$, $m_j = \max(a_{ij}, m_j)$; end;
                  end;
                  for $i = 1 \,..\, |V|$ do
                      for $j = 1 \,..\, |D|$ do $b_{ij} = \frac{a_{ij}}{m_j}$; end;
                  end;
graphing:         for all tolerance $t_i \in [u1, u2], step\ \delta$ do
                  Object function $ob = 0$;
                      for all $B_i \in B$ do
                          run modified global k-means over $b_i = \{b_{i1}, b_{i2}, \ldots, b_{i|V|}\}$;
                          generate cluster $C_{i1}, C_{i2}, \ldots, C_{il_i}$;
                          $ob = ob + \Sigma_{p=1}^{l_i} \Sigma_{q=1}^{|V|} ||C_{ip} - b_{iq}||$;
                      end;
                      plot $ob$ and $t_i$ in to objective function graph $G$;
                  end;
selection         $\hat{V} = V, \bar{V} = \Phi$;
                  for all $v_i \in V$ do
                      run the graphing process in $V - v_i$ over $B - b_i^T$
                      generate cluster $C_{i1'}, C_{i2'}, \ldots, C_{il'_i}$ and objective function $ob_i$;
                      plot $ob_i$ and $t_i$ in to objective function graph $G_i$ together with $G$;
                      if $ob_i > ob$ then $\hat{V} = \hat{V} v_i, \bar{V} = \bar{V} \cup v_i$; end
                  end;
evaluation        run graphing process based on $\hat{V}$ over $\hat{B}$ to
                          generate cluster $C' = C'_1, C'_2, \ldots, C'_{\hat{V}}$ and objective function $ob_{\hat{V}}$;
                          plot $ob_{\hat{V}}$ and $t_i$ as graph $G_{\hat{V}}$
                  if $ob > ob_{\hat{V}}$ then output $\hat{V}$ and $C'$;
                  otherwise
                          for all $v_i \in \bar{V}$ do
                  end;
end
```

Fig. 3. Feature selection algorithm using iteration; m_i is the maximin value of each feature; F represent all the feature extraction methods; u_1 and u_2 are the upper and lower boundary of t_i; $|\hat{B}| < |B|$ where $\hat{B}$ is the matrix using selected features only

The feature selection approach is shown in Figure 2. According to our approach, the feature collection and selection by means of an algorithm enclosing the chosen iteration algorithm - the modified global k-mean [2], [3].

Modified global k-means (MGKM) is an advanced version of the global k-means algorithm and k-means algorithm which is especially effective for solving clustering problems. Traditionally, k-mean algorithms randomly select instances as starting points to identify centroid of the given number of cluster. The computation is very expensive on large datasets. MGKM algorithm builds clusters starting from one centroid and iteratively adds one centroid at a time. An auxiliary cluster function is the key technique in MGKM which is minimized to find starting point for the next centroid. In particular, every starting point of new centroid (other than the first centroid) is detected using previous iterations.

Our algorithm conducts the search for an optimized feature subset using the modified global k-means for the evaluation of the current feature subset. It is run repeatedly on the phishing emails using various feature subsets and various tolerance values. The performance is estimated by objective functions using various feature subsets, where the subset with the lowest objective values is chosen as the iterated feature subset on which the induction algorithm runs.

Figure 3 illustrates the algorithm for selecting features. The algorithm can be divided into three components: graphing, comparison and evaluation.

1) We start with clustering emails D over all the possible feature set V collection. The modified global k-means algorithm generates clusters C and the clusters are estimated with an objective function ob value over a wide range of tolerances from 0.0001 to 0.1.
2) We then repeat the previous process with one feature less at a time, totally $|V|$ times. Again the modified global k-means algorithm generates clusters C_i where $i = 1, 2, \ldots, |V|$, and an objective function value ob_i for each feature removed. By comparing ob_i and ob we are able to identify features less important in clustering. Meanwhile, both informative features $\hat{V}$ and the possible insignificant feature sets $\bar{V}$ are identified, where the $\bar{V}$ contains a small number of features.
3) Finally, we estimate the selected features and identify

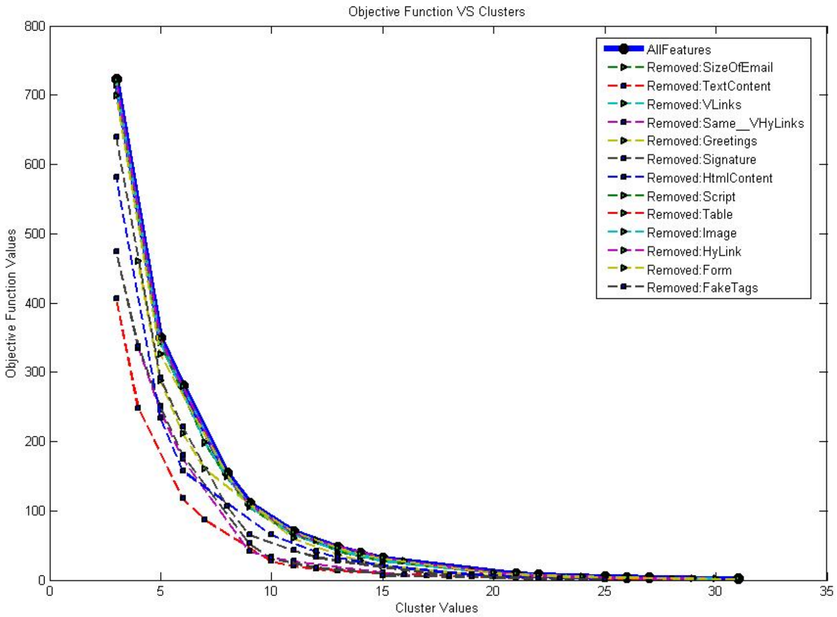


Fig. 4. Performance comparison of original feature set and each feature removed at once from it - values of objection function vs number of clusters

the optimal cluster using the comparison of the objective function value among different datasets. We apply the modified global k-means algorithm to $\hat{V}$ to generate the objective function value $ob_{|\hat{V}|}$. $\hat{V}$ is considered as the target vector if $ob_{|\hat{V}|}$ performs better than $ob_{|V|}$. Otherwise, we add features in $\bar{V}$ back to $\hat{V}$ to repeat the iteration process. One of the feature in $\bar{V}$ is added back to $\hat{V}$ at once in the first round, then any two features of $\bar{V}$ in the second round, until $|V| - 2$ features left. We generate the objective function value for each case, and identify the best feature sets $\hat{V}$ eventually.

The larger the dimensionality is, the bigger the value of the objective function should be. A feature is considered as a good discriminator when values of the objective function are approximately the same when the feature is included and excluded in clusters.

It is worth noting that the original feature set $|V|$ is not large, and the insignificant feature set $\bar{V}$ is rather small. Therefore, the process of step 3 is not time consuming.

V. Experimental Results

This section provides a comprehensive evaluation of text clustering and its feature selection. The data used in our experiment are the phishing emails collected from an Australian financial institution over five months. We have used a total of 2048 documents without any pre-defined knowledge about the documents.

We have fully implemented the algorithms described in Section III and Section IV. We used Java for document processing and MatLab for the learning process and plotting. The experiment is designed to illustrate the effectiveness and the reliability of the algorithm. The evaluation is designed as three major steps.

Step 1: Initial cluster generation

For the experiment, the emails were presented with a normalized matrix according to the description in Section III. Then the modified global k-means algorithm was executed using different tolerance values, ranging from 0.0001 to 0.1. For each different tolerance value, the different clusters and values of their objective functions are generated. Because the clusters and the values of the objective function are generated using all the features collected, we named them as initial clusters and initial values of the objective function.

Step 2: Feature elimination

We repeated the same procedure of Step 1 with a smaller feature set: each features is removed one at a time. For each run, we gained the same types of results as in Step 1.

The results of the experiments Step 1 and Step 2 are summarized in Figure 4 and 5 that plots values of the objective function vs the number of clusters shown in 4 and values of the objective function vs tolerance values shown in 5. The experimental results show the impact of that each individual feature causes.

Figure 4 describes the relationship between the values of the objective function over different numbers of clusters. The thick line represents the initial cluster, and the other dash lines represent another thirteen clusters when one feature was removed at a time from the original feature vector space. Figure 5 describes the relationship between the values of objection over different tolerance values. Notation are used

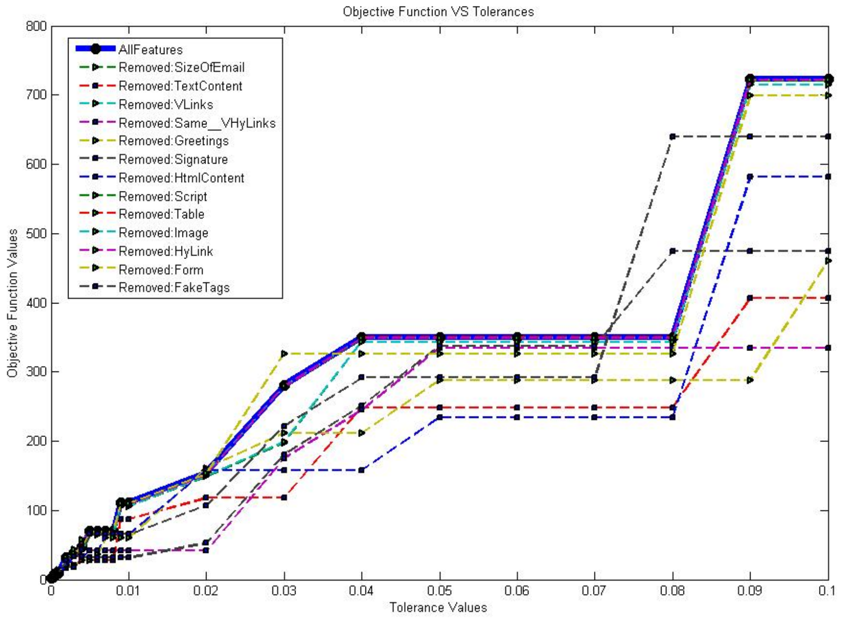
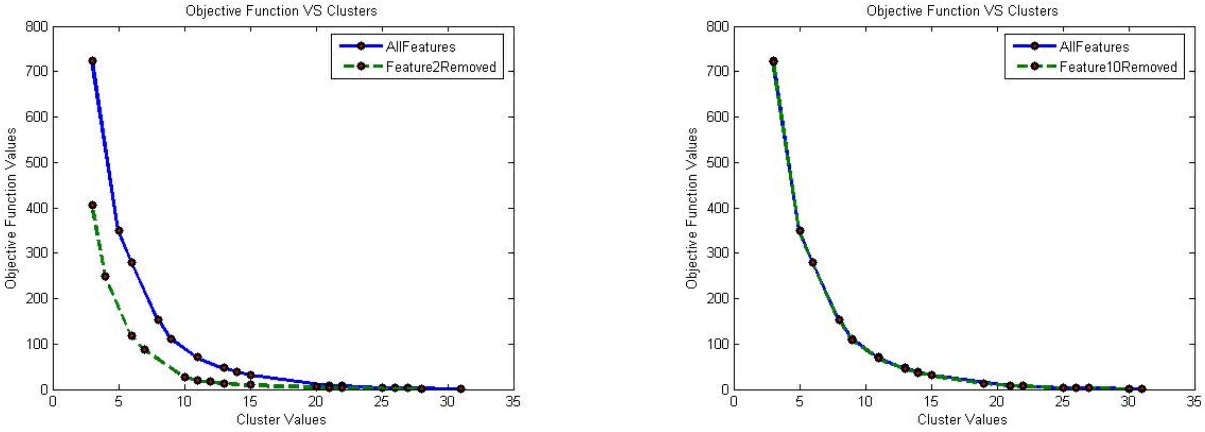


Fig. 5. Performance comparison of original feature set and each feature removed at once from it - values of the objective function vs tolerance values



(a) **(b)**

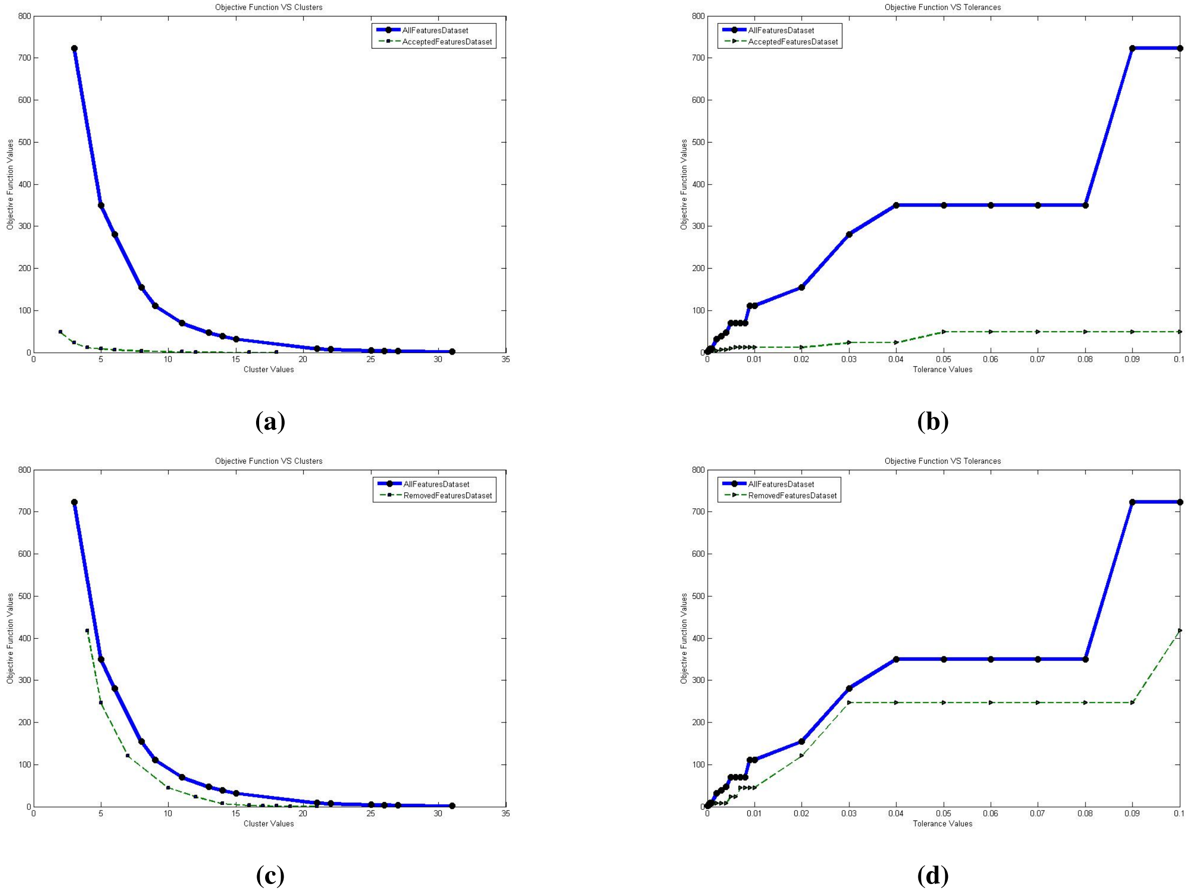
Fig. 6. Comparison of the values of objection function vs number of clusters (a) example of a bad feature (b) example of a good feature

here same as in 4 to denote the experiment over different feature spaces.

An interesting observation from Figure 4 and 5is that 4 gains lower objective function values when a feature is removed, at the same time 5 gains lower objective function values as well. The cluster and tolerance show strong support to each other so that the cluster and feature selection process are sound.

Figure 6 is a sub-figure of Figure 4 which contains the comparison of the original clusters and clusters with one feature removed. It is observed that with some removal of features (such as "Feature 2:text content", shown in 6 **(a)** the value of the objective function decreased considerably and more clear clusters are obtained. This indicates that these features like Feature 2 cause more noise and confusion for the clustering and they are considered as redundant features. Figure 6 **(b)** shows that after the removal of Feature 10 (number of image in an email), the value of the objective function remains the same as the original value which indicates that Feature 10 plays an important role in the clustering.

An investigation into the original dataset revealed that in most cases, Feature 2 has a zero value, therefore, removing such feature results in a lower value of the objective function. For another example, fake tags and its property are important features to identify a phishing email, since 89% from our data set the visible address are different from the invisible address to users.



(a) **(b)**

(c) **(d)**

Fig. 7. (a) and (b) are the performance comparison of selected features and original features; (c) and (d) are the performance comparison of removed features and original features in terms of values of objection function vs number of clusters and values of the objective function vs tolerance values

Step 3: Final cluster generation

Five features were considered as ambiguous in the feature selection process and were removed and eight features were selected. Clustering using the modified global k-means algorithm was applied on the selected features. Tolerance values from 0.0001 to 0.1 as before were taken into account and the clusters formed were recorded.

Figure 7 **(a)** and **(b)** show the values of the objective function based on selected features were substantially lower than on the original dataset. This indicates that more compact clusters were produced on the new dataset than on the original one.

We have also carried our experiment over the removed features as shown as in Figure 7 **(c)** and **(d)**. Though the dimensionality of this feature set is lower than the selected one, the values of the objective function **(c)** are much larger than the selected one **(a)**. **(a)** shows eleven clusters but **(c)** does not provide any good indication. The performance over the tolerance **(d)** behaves poorly compared to the performance in **(b)**. The figures here shows again that the feature selection process is comprehensive.

In all the cases, the tolerance and the values of the objective function are in a positive relationship. The tolerance graph 6 **(b)** shows the improvement of clustering on selected features. However, the objective function values and number of clusters are in a negative relationship. In the evaluation point of view, the lower the objective function values are relatively, the more efficient the clustering is. The graph helps us to determine a sharp trend which indicates a "treat-in" point of increasing clusters and decreasing values of the objective function, after which the change was almost minimal and consistent. This sharp bend corresponds to approximately 11 clusters. Hence, our preliminary investigation of results revealed that there will be 11 different clusters. Again, 6 **(a)** shows its support of having 11 clusters since after 11 clusters the values of the objective function are rather consistent. This leads us to the conclusion on the number of possible profiles from the structural analysis of the phishing emails.

Finally, the advantages clearly shown in the experimental results are (1) the efficiency of the automatic feature selection process (2) strong indication of correct clusters, and (3) tolerance and clusters support each other as expected.

VI. Conclusion

In this paper, we have presented a novel approach to cluster phishing emails using orthographic features for the purpose of provenance determination. The contribution of the work mainly consists of an evaluation method for the unsupervised task of clustering, usage of orthographic features and the learning algorithms.

Usage of orthographic features Most current information retrieval and categorized systems focus on text features. However, terms are largely similar in the same type of documents, therefore, they are not proper discriminators. On the other hand, orthographic features reflect the author's styles and habit, so that the features are more informative than text features in same type of documents. Experimental results carried out in this work show that orthographic features play an important role. In addition, low dimensionality is an advantage of orthographic features because we only need to increase the feature quality without considering the dimensionality reduction. Low dimension also guarantees a fast process.

Learning algorithm We have developed an algorithm to cluster documents and remove redundant features at the same time. We utilized the modified global k-means method repeatly over different datasets and identified redundant features. Experimental results show that a set of clear clusters are generated after the redundant features are removed compared with any other cases. Experimental evaluation on a large number of computations demonstrates that our clustering and feature selection techniques are highly effective and achieve reliable results.

Evaluation Methodologies Evaluation is always a challenging issue in text clustering because of the automation and lack of supervision. We have developed a comprehensive method to show the confidence for clusters generated with our algorithm. A number of experimental results show that the values of the objective function over the number of clusters and tolerance provide strong indications of the actual clusters. Simultaneously, the algorithm evaluates the goodness of using objective functions and proving that over a subset of good features.

While these results are impressive, there remain further possibilities for refining the approach. For example, the work presented in this paper focuses on content of emails because the content is the most complex part. It may be possible to analyse the headings of emails which are well structured to enhance the confidence of the clustering. We also intend to cluster phishing emails in another way: using the email structural presentation in terms of paragraphs, links, tables and their sequences. According to the preliminary investigation of results in this paper, there will be 11 phishing provenance, therefore studying the characteristics of each provenance will form the basis for future research.

One of the findings of Global Phishing Survey [13] by APWG (May 2009) is that phishers are increasingly using subdomain services to host and manage their phishing sites. We intend to parse and analyse URLs particularly Top-Level Domains (TLDs), subdomain, develop patterns to represent the URLs. Such patterns can be used in phishing detection and clustering.

References

[1] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26(2):1–29, 2008.

[2] A.M. Bagirov and K. Mardaneh. Modified global k-means algorithm for clustering in gene expression data sets. In *Proc. 2006 Workshop on Intelligent Systems for Bioinformatics (WISB 2006)*, volume 73, Hobart, Australia. CRPIT.

[3] Adil M. Bagirow and John Yearwood. A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems. *European Journal of Operatioanl Research*, 170:578–596, 2006.

[4] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *KDD-04, Processings of Internation Conference on Knowledge Discovery and Data Mining*, 2004.

[5] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Procedings of the Conference of Computational Learning Theory*, 1998.

[6] D. Boneh. Spoofguard. Technical report, http://crypto.stanford.edu/SpoofGuard/.

[7] Madhusudhanan Chandrasekaran, Krishnan Narayanan, and Shambhu Upadhyaya. Phishing email detection based on structural properties. In *Cyber Security Conference*, NYS, 2006.

[8] Manoranjan Dash, Kiseok Choi, Peter Scheuermann, and Huan Liu. Feature selection for clustering - a filter solution. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*. IEEE Computer Society, 2002.

[9] Manoranjan Dash and Huan Liu. Feature selection for clustering. In *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*. Springer-Verlag, 2000.

[10] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *Proceedings of the 16th International World Wide Web Conference (WWW 2007)*, May 2007.

[11] A. Hotho, S.Staab, and G. Stumme. Text clustering based on background knowledge. Technical report, University of Karlsrube, Institute AIFB, 2003.

[12] http://www.antiphishing.org/. Anti-phishing working group.

[13] http://www.apwg.org/reports/APWG_GlobalPhishingSurvey2H2008.pdf. Global phishing survey: Trends and domain name used in 2h2008, May 2009.

[14] http://www.gartner.com/it/page.jsp?id=565125. Agartner.

[15] Yifen Huang and Tom M. Mitchell. Text clustering with extended user feedback. In *Proceedings of International Conference on Information Retrieval, SIGIR'06*, pages 413–420, Seattle, USA, August 2006. ACM Press, NY, USA.

[16] Xiang Ji and Wei Xu. Document clustering with prior knowledge full text. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 405 – 412, Seattle, USA, August 2006. ACM Press, NY, USA.

[17] T. Joachims. Transductive inference for text classification using support vector machine. In *Proceeding of ICML-99*, 1999.

[18] Engin Kirda and Christopher Kruegel. Protecting users against phishing attacks. *The Computer Journal*, 2005.

[19] Christopher Kruegel, Giovanni Vigna, and William Robertson. A multimodel approach to the detection of web-based attacks, July 2005.

[20] Robert Layton and Paul Watters. Using differencing to increase distinctiveness for phishing website clustering. In *Proceedings of the Cybercrime and Trustworthy Computing Workshop (CTC-2009)*, Brisbane, Australia, 2009.

[21] Christian Ludl, Sean McAllister, Engin Kirda, and Christopher Kruegel. On the effectiveness of techniques to detect phishing sites. In *Proceedings of Detection of Intrusions and Malware and Vulnerability Assessment (DIMVA) 2007.*

[22] Liping Ma, Bahadorrezda Ofoghi, Paul Watters, and Simon Brown. Detecting phishing emails using hybrid features. In *Proceedings of the Cybercrime and Trustworthy Computing Workshop (CTC-2009)*, Brisbane, Australia, 2009.

[23] Stephen McCombie, Paul Watters, Alex Ng, and Brett Watson. Forensic characteristics of phishing - petty theft or organized crime? In *WEBIST*, pages 149–157, 2008.

[24] T. M. Mitchell. Machine learning, 1997.

[25] B. Ross, C. Jackson, N. Miyake, D. Boneh, and J. Mitchell. A browser plug-in solution to the unique password problem, 2005.

[26] B. Ross, C. Jackson, N. Miyake, D. Boneh, and J. Mitchell. Stronger password authentication using browser extensions. In *14th Usenix Security Symposium*, Baltimore, MD, USA, 2005.

[27] V. Roth and T. Lange. Feature selection in clustering problems. In *In Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA, 2004.

[28] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[29] Liu Wenyin, Guanglin Huang, Liu Xiaoyue, Zhang Min, and Xiaotie Deng. Detection of phishing webpages based on visual similarity.