Email Shape Analysis

Paul Sroufe[1], Santi Phithakkitnukoon[1], Ram Dantu[1], and João Cangussu[2]

[1] Department of Computer Science & Engineering,
University of North Texas Denton, Texas 76203
{prs0010,santi,rdantu}@unt.edu
[2] Department of Computer Science, University of Texas at Dallas
Richardson, Texas 75083
cangussu@utdallas.edu

Abstract. Email has become an integral part of everyday life. Without a second thought we receive bills, bank statements, and sales promotions all to our inbox. Each email has hidden features that can be extracted. In this paper, we present a new mechanism to characterize an email without using content or context called Email Shape Analysis. We explore the applications of the email shape by carrying out a case study; botnet detection and two possible applications: spam filtering, and social-context based finger printing. Our in-depth analysis of botnet detection leads to very high accuracy of tracing templates and spam campaigns. However, when it comes to spam filtering we do not propose new method but rather a complementing method to the already high accuracy Bayesian spam filter. We also look at its ability to classify individual senders in personal email inbox's.

1 Introduction

The behavior of email is something that is often overlooked. Email has been with us for so long that we begin to take it for granted. However, email may yet provide new techniques for classification systems. In this paper, we introduce the concept of email shape analysis and a few of its applications. Email shape analysis is a simple yet powerful method of classifying emails without the use of conventional email analysis techniques which rely on header information, hyperlink analysis, and natural language processing. It is a method of breaking down emails into a parameterized form for use with modeling techniques. In parameterized form the email is seen as a skeleton of its text and HTML body. The skeleton is used to draw a contouring shape, which is used for email shape analysis.

One of the largest threats facing the internet privacy and security of email users is spam email. According to the NY Times in March 2009, 94% of all email is spam. Email can contain malicious code and lewd content, both of which need to be avoided by 100%. The use of a behavior based detection method will

K. Kant et al. (Eds.): ICDCN 2010, LNCS 5935, pp. 18–29, 2010.

increase the accuracy and compliment current analysis methods in malicious and spam activity.

In this paper, we discuss a case study involving spam botnet detection. We also discuss the possible applications spam and ham filtering and social finger printing of senders. Recent papers presenting on this topic of botnet detection use network traffic behavior [1][2] and also domain name service blackhole listings [3], whereby botnets are discovered when they query the blackhole listings in DNS servers. By introducing shape analysis, one can further confirm the authenticity of the bot classifier.

The first application goes back to the proverbial spam question [4][5][6][7]. We look at the ability of shape analysis to correctly identify spam. In this study we are not trying to compete against the Bayesian filter, but rather compliment its decision process by offering non-content and non-context aware classification. The nature of the shape analysis classifier allows for both language independent and size independent email shape generation. This is believed to be very useful as the world becomes further integrated and spam comes in multiple languages to everyone.

In the second application, we look at the potential of email shape analysis to identify social context-based finger prints. We propose the ability to distinguish individual or group senders based on the social context. The data set for this study is one subject's personal email inbox.

The rest of the paper is organized as follows. The concept of the proposed Email Shape is described in section 2. Section 3 presents the case study email spam botnet detection. Section 4 discusses future work and their preliminary results on email spam filtering and social context-based finger print identification. Section 5 reviews some limitations of our study. Section 6 concludes the paper with a summary and an outlook on future work.

2 Email Shape

We define "shape" of an email as a shape that a human would perceive (*e.g.*, shape of a bottle). Without reading the content, shape of an email can be visualized as its contour envelope.

Email shape (e-shape) can be obtained from its "skeleton" that is simply a set of character counts for each line in the text and HTML code of email content. Let L denote the total number of lines in the email text and HTML code, and h_k denote the character count (this includes all characters and whitespace) in line k. A skeleton (H) of an email thus can be defined as follows.

$$H = \{h_1, h_2, h_3, ..., h_L\}\,. \tag{1}$$

Skeleton H can be treated as a random variable. Thereby the shape of an email can be derived from its skeleton by applying a Gaussian kernel density function (also known as Parzen window method) [8], which is a non-parametric approach for estimating probability density function (pdf) of a random variable and given by Eq. 2.

$$f(x) = \frac{1}{Lw} \sum_{k=1}^{L} K(\frac{x - h_k}{w}), \tag{2}$$

where $K(u)$ is the kernel function and w is the bandwidth or smoothing parameter. To select the optimal bandwidth, we use the AMISE optimal bandwidth selection based on Sheather Jones Solve-the-equation plug-in method [9]. Our kernel function is a widely used zero mean and unit variance given by Eq. 3

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}. \tag{3}$$

With this approach, an algorithm for finding e-shape can be constructed as shown in Alg. 1. Figure 1 illustrates the process of extracting e-shape. An example of four different e-shapes is illustrated in Fig. 2.

Algorithm 1. Email Shape
S = Email Shape(C)
Input: Email Text and HTML code (C)
Output: E-Shape (S)
1. FOR $i = 1$ to L /*L is the total number of lines in email HTML code */
2. h_i = character count of line i;
3. END FOR
4. $H = \{h_1, h_2, h_3, ..., h_L\}$; /* skeleton is extracted */
5. S =applying Gaussian kernel density function on$H$; /* e-shape is obtained */
6. Return S

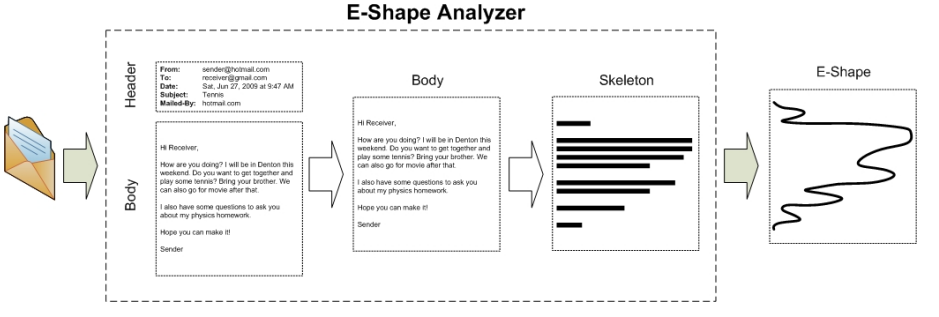


Fig. 1. Email shape analyzer

In summary, email shape is found by computing the number of character per line in an email. Almost every email has a text and HTML body. The lines are put into a file from which the Gaussian kernel density estimator smooths the rigid line graph into a normalized, smoothed graph. This graph is calculated for every email. We then performed a comparative function, called Hellinger distance, to find how closely each email shape is related.

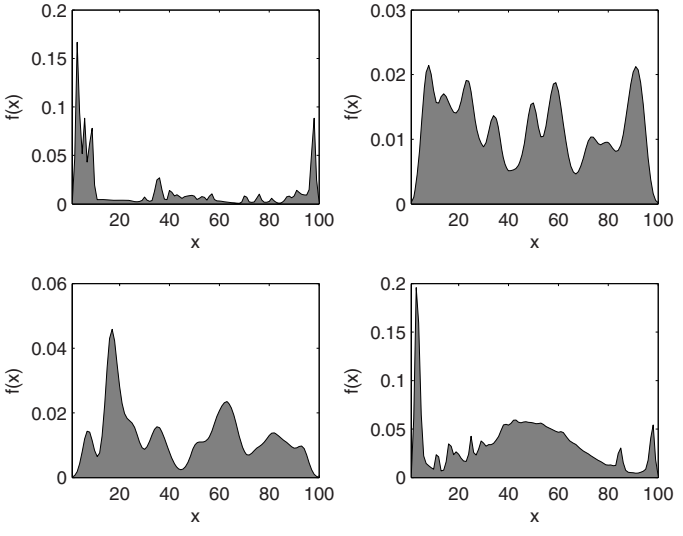


Fig. 2. An example of four different e-shapes

3 Applications of E-Shape

Our understanding of what shape analysis has to offer to the community is only at the beginning. We present, in the paper, a case study and two future work applications that outline some of the behaviors that shape can be used to analyze. First, is analysis of spamming botnets by template and/or campaign detection based on shape. By identifying similar shapes from different parts of the globe, one could surmise that they come from a matching bot host controller. (A bot is a compromised host that resides on the internet, usually without the host's controller's knowledge. The term bot has negative connotation and is usually associated with malicious behavior such as spamming, denial of service, or phishing. A botnet is a collection of two or more bots, and sometimes on the order of 10,000.) Second, spam filtering has become second nature to world. It has over 99% accuracy, but what of the last less than one percent? What were the content and context that were able to escape the filtering process? In this application we propose that e-shape analysis can be used to get closer to the goal of 100% spam classification. Third, e-shape analysis shows the discriminatory power to identify individuals on a personal level. In this application we build personal finger prints and turn our classifier over to the ham side of email.

3.1 Spam Botnet Detection

Spamming botnets are notoriously hard to pin point, often needing to use several methods to achieve decent accuracy. Here we present another tool to use in the

assessment of botnet detection. For this case study we gathered a data set of spam emails collected by Gmail's spam filter over the period of one month, during July 2008. The data set was over 1,100 emails in four different languages. The majority language was English. This data set was hand labeled into buckets based on content, size, and email type (e.g. Plain, HTML, Multipart). Each bucket would then contain similar emails, for example one group would contain emails sent that contained "Kings Watchmaker".

Hand labeling. To hand label thousands of emails we developed a program to display emails for ease of labeling. The program allows for a user to view a recorded history of previous labels, at any time refer to specific email for comparison, and resume previous labeling sessions. Files are written to an object text file, known as pickling, to preserve the email object format. The botnet label is written as a header directly into the email. A graphical user interface is included for the program.

After labeling several hundred of the emails, we started to see patterns emerge. We found evidence to support that botnet spammer's used templates to bypass spam filters, and they would fill in the blanks with the links and info they needed to get through (An example of the actual spam botnet template is shown in Fig. 3). The spam emails are very diverse, also shown by the multiple languages. The details of our data set is listed on Table 1.

Template Discussion. In the United States over 650 million email accounts are owned by four companies: Microsoft(MSN), Yahoo, Google and AOL [10]. Google comes in a distant third to MSN and Yahoo. They are very protective of their users and to get solicited emails to them can be an expensive process. We have evidence [11][12][13] to believe botnets are using specific templates to beat out spam filters. Seen in Fig. 3, a spammer would simply need to fill in the blanks and begin his campaign. The use of randomized or individually written emails for the purpose of spamming is not feasible on any small, medium or large scale campaign. It is of note to the authors that multiple botnets could be using the same template and be classified together. A separate method will be analyzed for distinguishing them in future work.

The total number of buckets from hand labeling was 52. For analysis we discarded buckets that had less than 10-emails per. This yielded 11 buckets. The shape of the testing email was derived using Alg. 1, then classified into different botnet groups. The measure of difference in shapes between these groups was based on Hellinger distance [14] since the e-shape is built with an estimated probability density functions (pdf). By using an estimated pdf, we are able to smooth out the shape from its rigid skeleton. It also normalizes the number of lines in the email, for use of Hellinger distance. The normalization of length is what provides a size independent way to calculate shape. Looking at the template in Fig. 3, a host spammer could add another paragraph with more links and not still not drastically change the normalized e-shape of itself.

Figure 4 shows two email shapes from a Chinese botnet. Figure 4(a) is larger than Fig. 4(b) by 22 lines, a difference of 11.8%. The two shapes are considerably similar and were mapped to the same bucket by the e-shape algorithm.

------=_NextPart_001_2D49_73AC2523.5E4E77CE
Content-Type: text/plain
Content-Transfer-Encoding: quoted-printable

Company Name
Motto Here
=20

Dear Name,

Run the erranking resultsRun the user-friendly and technology driven Tool P=
rogram.Try the FREE 90 day trial and tart achievingoutstanding search engin=
e placement and ranking results. Run the user-friendly and technology drive=
n Optimization Tool Program.Try the FREE 90 day trial and outstanding searc=
h engine placement and ranking resultsRun t he user-friendly and technology=
 driven Optimization Tool Program. Try the FREE 90 day trial and start achi=
evingoutstanding search e ngine placement and ranking resultsRun the user-f=
riendly and techn ology driven

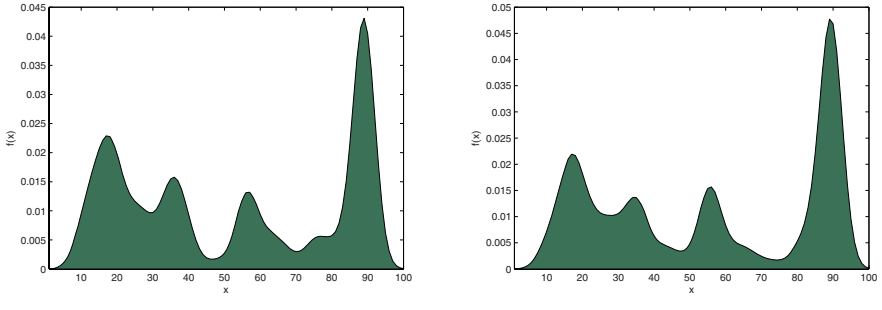
Sincerly,

John Smith
Manager Acounts
Company Software=20

Tel: your telephone
Fax: your fax
Web: your web site
=20

Copyright@Company Name.com

Fig. 3. An example of the actual spam botnet template



(a) Shape 1, 186 lines (b) Shape 2, 164 lines

Fig. 4. Showing size independence of shapes from the same botnet

The signature of each botnet group was computed as the expected value (mean) of the group. We used predefined threshold level at 0.08, which found to be the optimal threshold for our study. Hellinger distance is widely used for estimating a distance (difference) between two probability measures (*e.g.*, pdf, pmf). Hellinger distance between two probability measures A and B can be computed as follows.

$$d_H^2(A,B) = \frac{1}{2}\sum_{m=1}^{M}(\sqrt{a_m} - \sqrt{b_m})^2, \tag{4}$$

Table 1. Details of dataset for botnet detection experiment

Feature	Count
Total Emails	1,144
Email's Sizes of 1 to 100 lines	906
Email's Sizes of 101 to 200 lines	131
Email's Sizes of 201 to 300 lines	42
Email's Sizes of 301 to 400 lines	25
Email's Sizes of 401 to 500 lines	40
Emails in English	815
Emails in Chinese	270
Emails in Spanish	57
Emails in German	2

where A and B are M-tuple $\{a_1, a_2, a_3, ..., a_M\}$ and $\{b_1, b_2, b_3, ..., b_M\}$ respectively, and satisfy $a_m \geq 0, \sum_m a_m = 1, b_m \geq 0$, and $\sum_m b_m = 1$. Hellinger distance of 0 implies that $A = B$ whereas disjoint A and B yields the maximum distance of 1.

The accuracy of this data set is found from computing the number of correctly labeled emails in a bucket to the total number of emails in that bucket. A false positive indicates an email that was placed in the bucket but did not belong. A false negative would be the total number of emails, from hand labeling, that are in the rest of the buckets which belong to that bucket.

Figure 5 shows a promising accumulative accuracy rate of almost 81%. This number reflects the cumulative accuracy of all the buckets. While some buckets have a low accuracy, several of the buckets have a very good accuracy up to and including 100%, seen in Table 2. The evidence of a 100% accuracy bucket would show a positive match on an email campaign template. Accuracies below 50% are

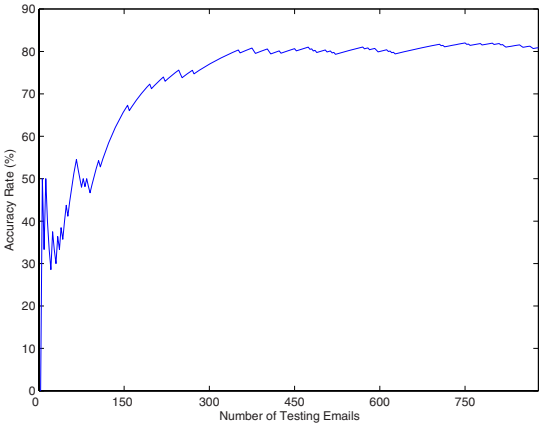


Fig. 5. A result of the botnet detection experiment based on 879 different size and language emails

Table 2. Accuracy rate of individual bucket

Bucket	Accuracy	False Negative	False Positive	Total Emails
1	41.37%	14	48	81
2	74.07%	20	20	76
3	80.95%	20	11	59
4	100%	0	0	129
5	68.75%	0	28	90
6	45.83%	0	25	67
7	100%	0	0	78
8	93.10%	14	6	81
9	88.00%	22	17	140
10	100%	0	0	118
11	100%	0	0	70

simply emails that are of similar shape. For example, bucket 6 is a mismatch of several botnet's from different languages and types of spam emails. Email shape analysis is showing good results in botnet and campaign classification, the purpose being to take context and specific content out of the classification process.

4 Future Work and Preliminary Results

In our on going work to discover and explore the full potential of e-shape analysis, we take a look at a couple of possible applications and also some preliminary analysis and results on them. Below we discuss the use of e-shape on spam filtering and on social context-based finger printing. In our finger print analysis we look at the capability of e-shape to differentiate senders from each other.

4.1 Spam Filtering

In this application of e-shape, we discuss the behavior that e-shape analysis can have on the spam filtering process. The Bayesian filter proves to be over 99% successful most all the time. However, to reach the goal of 100% further analysis is required. The Bayesian filter uses content and context to classify emails. The process could be enhanced using the method of shape analysis to "look at" if an email is spam or ham, taking content and context completely out of the equation. Surprise emails to the classifier that can't be categorized or are unique in manufacturing might make it through.

The data set used for this case study was the Trec 2007 corpus [15]. The Trec corpora are widely used in spam testing. The 2007 corpus was over 74,000 emails. However, for this study, only the first 7,500 emails were used for analysis. The corpus was approximately 67% spam and 33% ham and has been hand labeled by the Trec Team.

The method for comparison of spam versus ham was similar to that of botnet detection case study. Here we again used an unsupervised learning algorithm to classify data. We have developed a program that will take an email file in

MBOX format and calculate how many similar groups their are and classify in the same way as section 3.1. A testing email was classified to ham or spam based on the closest clustered group signature. The drawback of this process is that the buckets will need to be labeled by ham or spam, which is independent of our classifier. Once the bucket is known to be ham/spam future email's which are classified into the bucket will be labeled as such.

Preliminary results show an accumulative accuracy of about 70% for 7,500 emails. The accuracy is great considering that no content or context was even referenced. The ability for shape analysis to act as a spam filter would be recommended for use with emails that the Bayesian filter finds unsure about. Future work in that regard would be to implement shape analysis inside the Bayesian filter process.

4.2 Social Context-Based Finger Print Detection

This application is on using e-shape analysis to identify an individual's personal email finger print based on social context. We define personal fingerprint as the shape that one typically uses to contact others with. When an individual writes emails, it is believed that his shape will stay relatively the same, although length may change, the way he/she writes will not. An example of this would be an individual that creates a new line about every 40-50 characters versus a person that creates no new lines at all. It is also believed that this method can be used to reveal a user's clique's, as seen in [16]. A user will type differently to his/her boss and work mates than he/she would to their close friends. In this case study we follow the aggregate pattern of other users sending to a specific person.

For the data set, we used the top three different senders from one subject's inbox. The emails were collected over five months. Using e-shape analysis, we

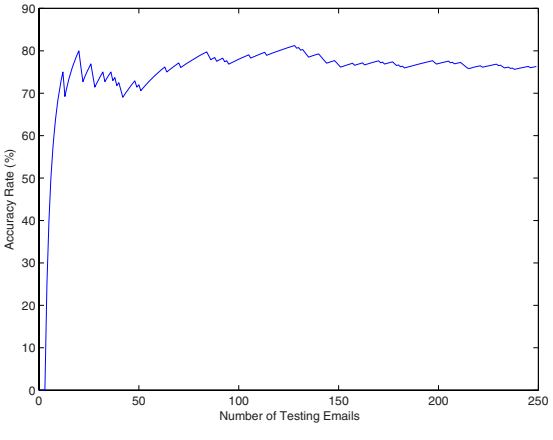


Fig. 6. A result of the social context-based individual's personal email finger print experiment based on three different individual email senders to a subject with total of about 250 emails

were able to distinguish these three senders to this subject, from unaltered emails (no thread deletion), with an accuracy of about 75% (see Fig. 6). The accuracy is considered good. Further refinement and post processing will be looked into in the future for better results. The current results now is using only the e-shape analysis.

Of the approximately 250 emails that are tested and of the three groups selected was a bi-weekly newsletter from a sales web site. The emails that came from this web site were classified with 100% accuracy and no false positives. The other two senders were from real human conversations.

This method reveals a very powerful tool in categorizing incoming emails when comparing non-human to human emails. Newsletters, advertisements, and solicitations can be moved separately by themselves to be reviewed later by a user, keeping priority emails displayed first.

5 Limitations of the Study

Currently the E-Shape analysis tool does not have a way to compliment its decision process by removing email threads and conversations. This drawback is reduced by the power of e-shape analysis, but it believed that we still yet have many abilities to unlock in this regard.

The shape analysis is a very useful tool to complement other tools as it can provide the deciding factor to many close decisions. Such is the example in spam detection where the content classifier can already achieve such a high accuracy. Some emails are short by nature, the ability for shape analysis to distinguish between others becomes limited. In the case of spam, short emails are common and the limitation impact of e-shape analysis will be mitigated to a large extent.

As mentioned earlier, and with any tool, the less information you give it, the less it can tell you. In the study of social context-based finger print detection, if a subject has a subset of friends that like to send web hyperlinks back and forth, the classifier will be unable to distinguish between users. Study of group based social awareness could be a possible application of this research.

Botnet detection is a challenging problem. There is not a singular solution to this threat, and combining the latest innovations only brings us a step closer. The purpose of E-Shape analysis for botnets is to bring the world one step closer. E-shape analysis is a tool capable of template/campaign identification to find a spamming bots before they are even able to send. Botnet identification is the next logical step of the process and can be supported with this tool.

6 Conclusion

In this paper, we present a novel concept of email shape (e-shape) and discuss three case studies using a hidden discriminatory power of e-shape. By using e-shape analysis we were able to detect botnet template/campaigns with about 81% accuracy. The botnet analysis can also be done with multiple languages and email sizes, which shows that the e-shape analysis is language and size

independent. Next, we discuss the capabilities of e-shape in spam filtering. Since e-shape is neither content nor context aware, it provides a unique point of view when looking at spam emails. We used the TREC 2007 corpus to test the spam filtering capabilities of e-shape. After running 7,500 emails through the email shape detector, we had a success rate of about 70%. Lastly, we looked at social context-based finger print detection, where we analyzed a single subject's email inbox. Using three different senders, we were able to achieve an accuracy rate of over 70%.

It is important to note that while the accuracy's of our system are not "high," the system of classification is taking content and context out of the classification process. This provides a very useful tool to complement existing methods and tools that currently handle emails, such as inching the Bayesian filter closer to 100% accuracy or assisting network behavior analyzers in determining botnet relationships.

As we evolve our understanding of what e-shape analysis can offer, we plan to improve the accuracies of the existing work and release more case studies. Currently the shape analysis routine does not have any smart way of handling email conversation threads or HTML code. This is the planned next direction of our work and is believed to offer a significant increase to ham labeling accuracy.

Acknowledgment

This work is supported by the National Science Foundation under grants CNS-0627754, CNS-0619871 and CNS-0551694.

References

1. Ramachandran, A., Feamster, N.: Understanding the network-level behavior of spammers. In: SIGCOMM 2006: Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications, pp. 291–302. ACM Press, New York (2006)
2. Strayer, W.T., Lapsley, D., Walsh, R., Livadas, C.: Botnet detection based on network behavior. In: Lee, W., Wang, C., Dagon, D. (eds.) Botnet Detection: Countering the Largest Security Threat. Springer, Heidelberg (2007)
3. Ramachandran, A., Feamster, N., Dagon, D.: Detecting botnet membership with dnsbl counterintelligence. In: Lee, W., Wang, C., Dagon, D. (eds.) Botnet Detection. Advances in Information Security, vol. 36, pp. 131–142. Springer, Heidelberg (2008)
4. Sinclair, S.: Adapting bayesian statistical spam filters to the server side. J. Comput. Small Coll. 19(5), 344–346 (2004)
5. Cormack, G.V.: Email spam filtering: A systematic review. Found. Trends Inf. Retr. 1(4), 335–455 (2007)
6. Cormack, G.V., Gómez Hidalgo, J.M., Sánz, E.P.: Spam filtering for short messages. In: CIKM 2007: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 313–320. ACM Press, New York (2007)

7. Wei, C.-P., Chen, H.-C., Cheng, T.-H.: Effective spam filtering: A single-class learning and ensemble approach. Decis. Support Syst. 45(3), 491–503 (2008)
8. Parzen, E.: On Estimation of a Probability Density Function and Mode. The Annals of Mathematical Statistics 33(3), 1065–1076 (1962)
9. Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society, Series B (53), 683–690 (1991)
10. Brownlow, M.: Email and webmail statistics (April 2008), http://www.email-marketing-reports.com/metrics/email-statistics.htm
11. Paul, R.: Researchers track Ron Paul spam back to Reactor botnet (December 2007), http://www.marshal8e6.com/trace/i/Template-Based-Spam,trace.996~.asp
12. Stewart, J.: Top Spam Botnets Exposed (April 2008), http://www.secureworks.com/research/threats/topbotnets/?threat=topbotnets
13. TRACElabs, Template Based Spam (May 2009), http://www.marshal8e6.com/trace/i/Template-Based-Spam,trace.996~.asp
14. Cam, L.L., Yang, G.L.: Asymptotics in Statistics: Some Basic Concepts. Springer, Heidelberg (2000)
15. Cormack, G.V., Lynam, T.R.: TREC 2007 Public Corpus (2007), http://plg.uwaterloo.ca/~gvcormac/treccorpus07/about.html
16. Stolfo, S.J., Hershkop, S., Hu, C.-W., Li, W.-J., Nimeskern, O., Wang, K.: Behavior-based modeling and its application to email analysis. ACM Trans. Internet Technol. 6(2), 187–221 (2006)