Evaluation of Spam Detection and Prevention Frameworks for Email and Image Spam - A State of Art

Pedram Hayati, Vidyasagar Potdar
Digital Ecosystems and Business Intelligence (DEBI) Institute
Curtin University of Technology
pedram.hayati@postgrad.curtin.edu.au v.potdar@curtin.edu.au

ABSTRACT
In recent years, online spam has become a major problem for the sustainability of the Internet. Excessive amounts of spam are not only reducing the quality of information available on the Internet but also creating concern amongst search engines and web users. This paper aims to analyse existing works in two different categories of spam domains - email spam and image spam to gain a deeper understanding of this problem. Future research directions are also presented in these spam domains.

Categories and Subject Descriptors

General Terms
Performance, Experimentation, and Security.

Keywords
Spam, Email Spam, Image spam

1. INTRODUCTION
Rapid adoption of the Internet in the day to day life of people has provided a platform for information generation and consumption. Internet is used on a daily basis to search for information and acquire knowledge. Several knowledge bases have been created on the Internet e.g. Wikipedia™ the online encyclopaedia. Wikipedia™ provides a good example of a more socialized Internet because the content within Wikipedia™ is collectively generated by its users, rather than webmasters or designated editors.

The ease with which content can be generated and published has also made it easier to create spam. Spam can be simply stated as information which does not add value to the web user e.g. inappropriate, unsolicited, repeated and irrelevant content in email messages, search results, blogs, forums, social communities and product reviews [4]. Some entities on the Internet often create synthetic content (or spam) to monetize it using online advertising like AdSense™ from Google™ [1]. In such cases the overall information quality is degraded drastically, which means when a user searches for a piece of information on the Internet, the likelihood of that information being spam is increased significantly [2]. This situation is getting worse day by day and hence research in spam prevention as well as detection is of prime importance to maintain quality of content on the web.

Hence the aim of this paper is to survey the current literature in the field of anti-spam with focus on specific anti-spam techniques used in email spam and image spam. The perspectives presented here is to show to the best of our knowledge how much work has been done in each of the spam domains and to highlight which domains require further investigation.

The rest of the paper is organized as following. Section 2 outlines the spammer’s motivations for spamming. Section 3 and Section 4 studies anti-spam techniques in email and image domain in detail. Section 5 provides an overall discussion, conclusion and future research directions.

2. SPAMMING MOTIVATIONS
In this section we outline the main motivations behind spamming based on their importance. To answer this question we group the main motivations into five different categories:

1. Revenue Generation: This is the main motivations behind web spam. Revenue can be generated via publishing advertisements on websites. Google AdSense™ program makes it very easy for publish ads on website and generate revenue [1]. Spammers exploit this service by generating synthetic content (or copied content) and then monetize it using AdSense™ [1].

2. Higher Search Engine Ranking: The next logical motivation for spammers is to incorporate search engine optimization techniques to get their website a higher rank in search results, which results in more traffic and consequently

1 https://www.google.com/adsense/
more revenue via advertising [1]. Hence this is the second most important motivation.

3. **Promoting Products and Services**: This can be considered as a subset of revenue generation. In this case spammers are paid by companies to promote their products or services [1]. Spammers often resort to emails, forums, review websites and blog comments to channel this information.

The above three motivations are passive since they do not intrude user privacy or security; however the next two motivations directly affect user’s security on the Internet.

4. **Stealing Information**: Here the main motivation of the spammer is to setup hidden programs in user’s computers to gain back door entry [3]. These programs run on users’ computers and perform some malicious actions like popping up an advertisement, opening websites, and stealing user information such as email address to send email spam or can even tap key logs.

5. **Phishing**: Here the main motivation is to steal sensitive information (such as credit card numbers, password, etc.) by masquerading as a legitimate identity.

Given above five spamming motivation categories, we found that revenue generation, promoting products & services, stealing information, and Phishing are the main motivation behind email and images spam.

In the next sections we study anti-spam methods in email and image spam domains to give the reader an in depth understanding of the current state of the art in anti-spam.

3. **EMAIL SPAM**

Email spam refers to sending irrelevant, inappropriate and unsolicited email messages to numerous people [4]. This is possible due to the low entrance barrier and low cost of sending emails, which makes it one of the most popular forms of spam [4-5]. The purpose of email spam is advertising, promotion, and spreading backdoors or malicious programs. Currently Phishing is also considered as one of the main goals of spammers when employing email spams. Figure 1 shows an email spam message in different language which shows that spammers nowadays send more customized email spams.

![Figure 1 – Screenshot of Email Spam in Arabic Language](image)

To tackle the problem of email spam many methods are presented, some of which are being used [6-10]. These methods use either detection or prevention strategies to combat spam. Table 1 outlines the latest researches in the field of email spam and in section 3.1 we critically evaluate them.

<table>
<thead>
<tr>
<th>Theory or Heuristics</th>
<th>List of features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HoneySpam, 2005 [10]</strong></td>
<td>The idea presented with HoneySpam is to combat spammers at the source instead of the destination which saves bandwidth and reduces traffic. HoneySpam is derived from HoneyPots (a machine or system which entices attacker by pretending as a real victim for gathering as much information as possible for future defense). HoneySpam builds dynamic web-pages which are linked together and contains huge amount of special valid email accounts to speed down email harvesting (the act of going through webpage in order to find new email addresses to send spam email). Each of these special email accounts are handled by HoneySpam SMTP (Simple Mail Transfer Protocol) server which track spammer activity. Additionally, HoneySpam provide fake open proxy or relays (a type of Internet service which is used for being anonymous) to log spammers’ activity, save network traffic, and block them.</td>
</tr>
<tr>
<td><strong>Phonetic String Matching, 2006 [6]</strong></td>
<td><strong>⇒</strong> Body of message <strong>⇒</strong> Email headers <strong>⇒</strong> Mapping graphical notation into original character. E.g. 'V' maps into 'v'), <strong>null-char removal</strong> (The procedure of removing useless non-alphabetic characters. E.g. ' '), <strong>key-specific disambiguation</strong> (The procedure of finding analogous letters in place of remaining non-alphabetic characters, similar to normalization procedure but applied after null-char removal stage), and <strong>phonetic transcription</strong> (The procedure of looking for corresponding arrangement of phonetic symbols represent similar English pronunciation e.g Buy and Bye). The result of each task is sent to approximate string matching (The procedure of insertion, deletion and swapping of adjacent characters which returns most likely related strings to original string) that results in 4 flags. These 4 flags combined in rule composition module (The procedure of flags eventually combining) to mark message as spam or ham.</td>
</tr>
</tbody>
</table>
3.1. Critical Evaluation

This section expands on the previous section and provides a critical evaluation of the current email anti-spam techniques along 8 different criteria (as shown in Table 2), by attempting to answer the following questions:

1. Is the approach a prevention or detection technique?
2. Can the technique be applied to detect non-English language spam?
3. Is the method content based method or meta-data based?
4. Does it take time to study the behaviour and build individual user profiles?
5. Can this technique detect spam on the fly?
6. Does it prevent spammers from consuming network bandwidth? e.g. honeypots, blacklists
7. Is the technique scalable enough to be implemented for a public email server?
8. Does the technique use supervised or unsupervised machine learning approach?

The notable answers to some of these questions will now be discussed in detail for selected frameworks. There are many types of techniques used to identify email spam, but majority of these techniques are detection based [6-8] since prevention techniques [4, 10] are difficult to implement in emails.

When research in this area began one of the first techniques was to block email selectively from certain domains and IP addresses and prevent further relay of email from unauthorized domains [4]. However the main drawback of this approach was that it was quite difficult to maintain such black lists as the number of SMTP servers started increasing at a rapid pace.

Hence researchers tried to look for automated ways to manage spam and hence content based spam detection methods came into picture. Several methods were proposed that used Naïve Bayesian filtering [11], memory based approaches [12], keyword based filtering [13], genetic programming [14] and trust rank [15]. Later statistical spam detection techniques were proposed as in [16], which were language independent.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevention (P) or</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>P</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>Detection (D)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strategy?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applicable to Non-English Language?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Possible</td>
<td>Possibl e</td>
<td>No</td>
<td>Possibl e</td>
</tr>
<tr>
<td>------------------------------------</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----------</td>
<td>----------</td>
<td>----</td>
<td>----------</td>
</tr>
<tr>
<td>Content based (C) or Meta-Data based (M)? Or both</td>
<td>C</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>-</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Does it need time to build user profile?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Detect spam on the fly?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Prevent spammers from consuming network bandwidth?</td>
<td>No</td>
<td>Possible</td>
<td>No</td>
<td>Possible</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Scalable to implement on a public email server?</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Supervised (S) or Unsupervised (U)?</td>
<td>S</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

Recently five other techniques have been proposed in the literature, which we have outlined in Table 1. We can categorize these techniques as content based and meta-content based techniques.

The phonetic string matching approach [6] as mentioned in Table 1 analyzes content of the email using 4 filters and flags it as spam or ham. One of the main drawbacks of this approach is that it can only function for English language, since the phonetic transcription module (PTM) is only designed for English. Secondly there is a potential flaw with PTM and it could give false negatives if a genuine user’s level of English is poor, specifically people from non-English background nations. They may spell ‘Buy’ as ‘BYE’ although they really want to mean ‘Buy’. Same applies to the normalization module and null-character removal. For example, in English language underlines do not signify anything important, however in other languages like Farsi or Arabic it is used to emphasize a particular word. As shown in Figure 1, spam is not only restricted to English language but it can also be seen in other languages like Arabic. Hence the way in which this method can be improved would be by developing language specific normalization and related filters.

Compared to [6], the ProMail system [7] has a distinct advantage i.e. it is not content dependent and hence can be applicable to non English language spam. ProMail operates on the meta-data by looking at some features like email address, IP address etc. and developing a user profile. However the trade-off with this method is that it takes time to develop this profile and every time a new user enters the system it would take some time to study its behaviour and develop the user profile. Hence this approach cannot be used on the fly and spammers can manipulate their behaviour to game the system easily i.e. they can use proxies to hide their IP address to bypass the spam filter.

Another similar approach presented in [9] uses SMTP logs to analyse user profiles. Here they use similar features but derive composite equations or ratio’s e.g. communication reciprocity, which looks at the percentage of interactive neighbours. These advanced equations can be implemented in a private email network where users can only have one email account. However in public email services like Yahoo™ or GMail™, this technique cannot be implemented successfully since the spammer can easily bypass these filters by creating another genuine account and interacting like a genuine user. Another downside of this approach is that it can take long to generate these profiles and these profiles can also be very huge and occupy lot of space on the server. The authors also do not state how long these profiles should be archived. If there are fewer profiles to monitor the spam filter can be optimized to run faster. In addition, some commercial free email services like Yahoo™ and GMail™ these days uses CAPTCHA to prevent email spam, but spammers would normally not use such email services. Hence the need to develop a robust system becomes an open research question.

Table 2 evaluates these techniques along eight dimensions as mentioned earlier. As shown in Table 2, two of these methods are based on prevention strategy while eight other methods are detection based. Five of those can be used in detecting spam that is not in English language, like [4, 7-10]. These techniques use meta-data to detect spam, which could be IP address, SMTP logs, email header information etc. However some of these techniques like [12, 13] may be enhanced to operate in non-English languages, but at this stage it is not implemented and since they are content based techniques they would require language specific filters to classify spam. Those techniques that are applicable to operate on English language use language specific filters like phonetic string matching [6], or keyword matching [13]. Compared to content based approaches, meta-data based methods create user profiles to combat spam and are applicable to non-English language [7-9] except [4], which uses blacklisting approach to prevent spam. However meta-data based methods [7-9] can detect spam if they have spammers’ profile in their training data set. On the other hand content based methods cannot detect spam on the fly since it is necessary for them to go through email content [6, 11-14]. Additionally, methods like [8] uses another approach to detect spam on the fly and that is comparing SMTP header with SMTP RFC protocol. Other than that we also found that most of the proposed methods in the email spam domain do not prevent spammers from using sensitive network bandwidth, however [4] they limit bandwidth by blacklisting. Apart from this, techniques like Promail and SMTP Log [7, 9] are creating user profiles to
detect spam, but based upon user profile interconnection it might be difficult to implement them in public email services. Finally we also found that most of methods use supervised learning which require manual intervention and can be a maintenance issue compared to other 3 unsupervised methods [7-9]. We now review image spam methods.

<table>
<thead>
<tr>
<th>Theory or Heuristics</th>
<th>List of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual feature and near duplicate image spam detection, 2008 [19]</td>
<td>➔ Pixel coordinate ➔ Colour attribute (Average RGB, Colour Histogram, Colour Moments) ➔ Texture attribute (Autocorrelation, Edge Frequency) ➔ Shape Features (Geometric Movement, Eccentricity)</td>
</tr>
<tr>
<td>Image Spam Hunter, 2008 [21]</td>
<td>➔ Colour orientation histogram ➔ Gradient orientation histogram</td>
</tr>
</tbody>
</table>

4. IMAGE SPAM

To get rid of anti-spam filters in email spam currently some spammers put their spam content into the images (i.e. they embed text such as advertisement text in the images) and attach these images to emails (Figure 2). Those anti-spam filters that analyse content of email cannot detect spam text in images [17].

4.1. Critical Evaluation

In section 4 we covered some of the latest image anti-spam detection techniques from the literature. This was done to identify study and extract the different features used to classify spam vs. ham content. We present a critical evaluation of the anti-spam techniques surveyed along 5 different dimensions (as shown in Table 4), by attempting to answer the following questions:
1. Is the technique content based or meta-content based or uses both approaches?
2. If it is content based, is OCR technique used for assessment?
3. Is the technique fast enough to scale on public email servers?
4. Does the technique use supervised or unsupervised machine learning approach?
5. Is the technique robust against content obscuring attacks?

The noteworthy answers to some of these questions will now be discussed in detail for selected techniques.

According to our survey we discovered that Fast Image Spam Classifier (FISC) method [17] is reasonably fast in detecting key features like average colour, colour saturation etc., however clever spammers can easily exploit these features and create synthetic images, which would not be flagged by spam filters. For instance, the image size or image format can be easily manipulated or the average colour or colour saturation could be adjusted.

In comparison to FISC [17], near-duplicate image spam (NDIS) method is relatively robust compared to other methods like OCR or object recognition techniques etc [23]. Nevertheless, the basic foundation of this method is not practical and realistic because the authors postulate that spammers always use a similar template when sending image spam emails, but this can be considered a very weak assumption to begin with.

A more recent method [19] enhances NDIS by incorporating some new features, which the authors assume offer much better detection capability along with using some ideas from FISC [17, 19]. But this method is also flawed since it builds up from a weak foundation as stated above. Compared to these three method [17, 22, 23] the method proposed in the Image Spam Hunter utilizes probabilistic boosted tree, which is much faster than SVM used in earlier methods [17, 18]. However, the authors themselves point out that some of the weaknesses of their method includes inaccurate detection if image contains scanned documents or real life photos.

Other than these papers [17-19,21] we also investigated seven other papers [22, 24, 29] as shown in Table 4. From the 11 image anti-spam methods, majority are content based [18, 21, 22, 24-27, 29], one of them is meta-data based [28] and one is both content and meta-data based [17]. Content-based methods rely on using OCR to assess content, which is time consuming and CPU intensive [22, 24, 29], hence recent methods do not apply it anymore and they use other image attributes to classify spam. Additionally, OCR based techniques are reasonably slow to implement in public email services [22, 24, 29]. Furthermore we found that almost all current image anti-spam methods are detection based rather than prevention based. In addition to this, we also found that most of the methods outlined in the literature are robust to content obscuring except [22, 24, 29].

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Content based (C) OR Meta-data based (M) OR both?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Fast enough to scale on public email servers?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Supervised (S) or unsupervised (U)?</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>Robust against content obscuring attacks?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

5. DISCUSSION AND CONCLUSION

In this paper we investigated email spam and image spam methods. Each of these domains employs specific spam detection and prevention techniques. First of all, email spam has been the earliest and largest contributor to spam. As mentioned earlier email spam started when email services began in early 1990s. Over the years spammers have become smarter and more diligent and have introduced image spam as a technique to by pass the most advanced email spam filters. The textual content of the email is now transmitted via images.

Detecting image spam is a difficult job however we surveyed some techniques that do offer some relief in this case. Some methods are using OCR based techniques while others rely on image attributes like colour histogram etc.

The main difficulty with image spam is that it is processor intensive. To analyse an image for spam detection takes more resources and computing power than simple text. From the spammer’s perspective, image spam is a more difficult task, but can make the process faster by automating the process to convert text to images. A recent paper also outlines how PDF documents are now used for spamming [30].

Based on our study we identified several research directions, which can be pursued in the field of email and image spam detection. We briefly outline them here.

The spam detection approach using the log from SMTP servers cannot be implemented in a public email server.
because the spammers create fake accounts and try to send emails to each other and try to imitate the behaviour of a genuine user. To solve this problem is still an open issue that needs to be investigated by researchers looking into email spam.

Some techniques generate user profiles for detecting spam emails; however the time it takes to develop such profiles can be much longer than the life of that email account. Spammers often use new email accounts hence user profiles should be created in a short time frame and should be disposed off once that account has become inactive, to improve the overall filtering process. Currently not a lot of work has been done in this direction and is a emerging and challenging filed of research.

It would also be interesting to see new research in the areas of content based spam detection for spam in non-English languages. As shown in Figure 1 we can see that spammers are now starting to use languages other than English to send spam. This particularly becomes a major issue with respect to spam in Mandarin language which is the main language for China (which has a population of around 1.3 billion people). There has been some work done in this field however it is still an interesting area of research [31, 32]. For instance, spam filters for “Hindi” currently do not exist; however it would not be long before spammers start targeting users with Hindi, which is spoken by majority of Indian population. This problem of lingual spam would become even more difficult when image spam is used.

In the context of image spam most of the statistical methods rely on identifying some kind of noise introduced by the spammers when spamming, however if the spammer start using real world pictures, it would be very hard to detect potential spam activity. This can be successfully done by lowering the ratio of spam content vs. original image content. However under bandwidth constraints the use of real world images can be an issue, since the file size is much bigger than a normal image with few colours.

At the moment OCR technique are slow and many languages do not have robust OCR techniques as superior as those available in English. So if spammers start using image spam in a different language detecting that would be difficult using OCR. Currently OCR techniques are rated lower because it takes too much processing power but in the future this scenario may change and hence it is not advisable to completely neglect spam detection research using OCR.

Other than that most of the surveyed techniques are detection based, hence the potential for research in spam prevention techniques is much more interesting. Currently newer versions of CAPTCHA and challenge-response based methods are being developed.

Spam techniques evolve over time as we have seen in the case of email spam. When email spam became difficult, spammers began using images to bypass the email spam filter. We assume that this evolution will continue and would spread across other spam domains. For instance, currently splogs are filtered based on their content and some other statistics, however if spammers begin to use genuine content from a blog with high page rank, and add images, which contain spam, it would be difficult for search engines to differentiate spam vs. ham. The option of using existing image spam detection methods is an alternative, however image spam detection techniques are itself in its infancy and a lot of work is required in this domain too.

In the future we could also see spammers targeting image search engines, where they would attempt to index spam images in the search engines database. This can be easily achieved by using fake anchor text, since majority of search engines currently rely on anchor text to predict the image content.

Last but not the least, what would happen if spammers start using well known, widely accepted security technique i.e. encryption or public key. Normally many people publish their public keys, which can be used for sending private emails; spammers can use those public keys to send spam emails, in which case it would be impossible to filter the content since it is encrypted.

REFERENCES


