FEATURE

THE MEDIUM OR THE MESSAGE? DEALING WITH IMAGE SPAM

Catalin Alexandru Cosoi
BitDefender, Romania

Internet users have recently become used to a new category of spam message: one in which the content is delivered in the form of an image attachment. A year ago, such 'image spam' accounted for approximately 10% of the total amount of spam circulating. In recent months, however, spammers have noticed that many of the current anti-spam solutions are almost ineffective against this trick so they have started attacking this niche in earnest. Image spam has increased to 30–40% of the total amount of circulating spam, with the addition of random noise making almost every image unique. Detection rates have dropped even further.

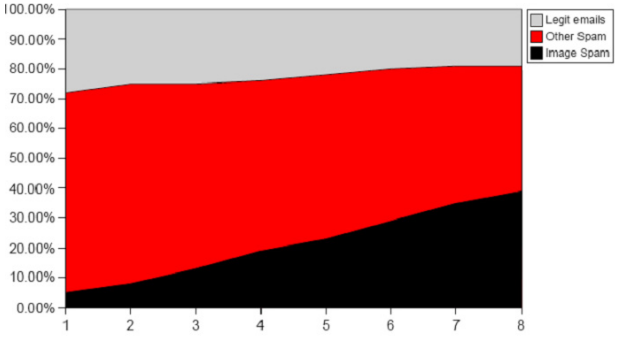

Figure 1: The evolution of image spam during the months March – October 2006. (Months labelled 1 – 8.)

Performing any sort of content analysis on such emails requires an Optical Character Recognition (OCR) module. Yet common OCR filters are computationally expensive and their accuracy leaves much to be desired.

An alternative approach would be to use a filter that ignores the text contained within the image (i.e. the message, from a human point of view) and instead learns by experience some common characteristics of the images themselves (the chosen medium or vector of communication), thus achieving reliable detection. With this idea in mind, our research concentrated on finding a way to represent and subsequently detect specific characteristics of spam images.

We were looking for a comparison function that is permissive enough to ignore noise, but sufficiently discriminative to avoid false positives, while still demanding reasonable amounts of processing power.

For this purpose, we took a new look at histogram extraction and comparison (a histogram can be defined in this case as a list of colours and their relative preponderance in an image; it tells us what colours exist in an image and how many pixels are of a given colour).

These are, of course, tried and tested techniques for content-based image retrieval (CBIR) that are fast enough to be run on almost any modern system, and also discriminating enough to provide a pretty good detection rate.

However, the false positive rate of such techniques is rather high. Of course, when you're just searching for pretty pictures, more *is* better. In an anti-spam solution, on the other hand, false positives are a serious problem. The types of distance functions commonly met in old CBIR systems are: histogram Euclidean distance, histogram intersection distance and histogram quadratic (cross) distance. Of all these options, the most appealing for an anti-spam engine would be the histogram intersection distance.

Experimentation revealed that this is potentially useful because it can ignore changes in background colours, but problematic when changes appear in the foreground. Also, if the noise added by the spammers uses the same colours but in a different quantity, these colours will add to the distance.

With a view to alleviating the problems of the classical algorithm, we introduced a new type of histogram distance, designed specifically for spam images – let's call it Spam Image Distance, or SID.

Equation 1 shows the definition of SID, where a, b and c are the quantities of red, green and blue, and h(a, b, c) represents the number of pixels occupied by the colour created by mixing a% red, b% green and c% blue.

In other words, we consider not only identical colours, but also similar ones, which can differ from the original by no more than δ values, with the restriction that we consider this element in the sum to be non-zero only if the difference between the sizes of the bins is smaller than Δ. These two parameters can be determined using Receiver Operating Characteristic (ROC) curves (simultaneous comparison of sensitivity and specificity), trial and error, or by using a machine-learning technique.

While this new technique can be shown to perform well on 'clean' images, there remains the problem of noise

$$SID(h,g)=\frac{\left(\sum_{A}\sum_{B}\sum_{C}\min\left(h(a,b,c),g(a\pm\delta,b\pm\delta,c\pm\delta)\right)\left|\frac{h(a,b,c)-g(a\pm\delta,b\pm\delta,c\pm\delta)}{\min\left(|h(a,b,c)|,|g(a\pm\delta,b\pm\delta,c\pm\delta)|\right)}\times 100\le\Delta\right)}{\min\left(|h|,|g|\right)}$$

Equation 1.

elimination. Fortunately, the techniques used by spammers to add noise or otherwise obfuscate the images are well known to us.

Common 'noising' techniques catalogued at this time include:

a. Adding random pixels to the image.

b. Animated GIFs with noisy bogus frames.

c. Similar colours between different parts of the text in the image.

d. A long line at the end of the image (some kind of border) with random parts missing.

e. Splitting the image into sub images and using the table facilities in HTML to reconstruct the image.

f. Sending different sizes of the same image.

g. Image poisoning – inserting legitimate image content such as company logos into spam messages.

The arsenal of countermeasures is similarly wide. For instance, to eliminate random pixel noise from an image histogram we can use this simple function:

$$H' = \left\{ c_i \in H \left| \frac{|c_i|}{\sum_{j=1}^{|H|} |c_j|} \times 100 > \gamma \right. \right\}$$

and let SID deal with the outcome. Another useful trick is to 'stitch together' the histograms of images embedded in HTML tables (if they are sufficiently similar) and then let SID consider the resulting composite histogram.

The distances found by the SID function are used to compare images that are already in the spam database with images that we want to add. If the image analysis returns a score smaller than a threshold T, then we add the image. Otherwise, we consider it to be a known image. The SID can only fail if an image is entirely new or if it is a malformed image from which we can't extract a histogram. In our current experience, new images appear at the rate of one per day.

Some care should be applied to deciding exactly what the filter learns, as spammers have started using company logos as noise in spam images. Misidentifying a company logo as a spam image could create a serious problem.

Some other sources of false positives exist as well. Using a filter that compares histograms does not tell one anything about the content of the pictures (the colours of human skin, for instance, are the same no matter whether the picture is explicit in nature or just an innocuous vacation snapshot).

DETECTION RATES

When run against the *BitDefender* corpus of spam images (a few million samples extracted from real spam) SID shows a 98.7% detection rate. Within the corpus 1.23% of images are malformed and we can't extract the histograms for those pieces of spam, but this is not considered to be a significant problem since the image cannot be seen by the user either.

A further 0.03% represent false positive results. If we delete from the corpus all those images that are malformed, the detection rate quickly jumps to 100%.

We believe that SID is a worthwhile addition to the arsenal of any modern spam hunter and that advances in noise reduction will further improve the potential of this already very useful tool.

BIBLIOGRAPHY

[1] Sablak, S.; Boult, T. E. Multilevel color histogram representation of color images by peaks for omni-camera. Proceedings of IASTED International Conference on Signal and Image Processing, 1999. See http://vast.uccs.edu/~tboult/PAPERS/IASTED99-Multilevel-color-histogram-representation-of-color-images-by-peaks-for-omni-camera--Sablak-Boult.pdf.

[2] Stevens, M.R.; Culberston, B.; Malzbender T. A histogram-based color consistency test for voxel coloring. Proceedings of 16th International Conference on Pattern Recognition, 2002. See http://www.hpl.hp.com/research/mmsl/publications/vision/icpr02-final.pdf.

[3] Scheunders, P. A comparison of clustering algorithms applied to color image quantization. Pattern Recognition Letters Vol. 18, No. 11, 1997. See http://webhost.ua.ac.be/visielab/papers/scheun/prl97.pdf.

[4] Pass, G.; Zabih R. Histogram refinement for content-based image retrieval. Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision, 1996.

[5] Ling, H.; Okada, K. Diffusion distance for histogram comparison. IEEE Conference on Computer Vision and Pattern Recognition, 2006. See http://www.cs.umd.edu/~hbling/Research/Publication/336_ling_h.pdf.

[6] Graham-Cumming, J. The rise and rise of image-based spam. Virus Bulletin Spam Supplement. November 2006 p. S2. See http://www.virusbtn.com/sba/2006/11/sb200611-image.