
Using visual and semantic features for anti-spam filters

F.Gargiulo, A. Penta, A. Picariello and C. Sansone
Dipartimento di Informatica e Sistemistica
Università di Napoli “Federico” - Napoli - Italy
{francesco.grg, a.penta, picus, carlosan}@unina.it

Abstract

It is well known that Unsolicited Commercial Emails (UCE), commonly known as *spam*, are becoming a serious problem for email accounts of single users, small companies and large institutions. The presence of spam can seriously compromise normal user activities, forcing to navigate through mailboxes to find the - relatively few - interesting emails, so wasting time and bandwidth, occupying their storage space. Eventually, they have often unsuitable content (as a pornographic material advertising) that could be illegal for the minors. In this realm, different countermeasures to spam have been proposed, using *regulatory* or *technical approach*. The legislative approach doesn't obtain the desired results. A variety of technical approaches are thus implemented in different anti-spam filters currently used to detect the spam content [8]. In the past, researchers have addressed this problem as a text classification or categorization problem. However, as spammers' techniques continue to evolve and the genre of email content becomes more and more diverse, keywords-based anti-spam approaches alone are no longer sufficient.

Different techniques are used to analyze the mail text, the majority are learning-based approach. Considering the spam detection as a binary classification problem, several algorithms from learning theory field can be used, such as bayesian algorithms [5] or Support Vector Machine (SVM) [6]. These systems, using the acquired knowledge, are able to discriminate the synthetic features in order to reject the mail considered as spam. Note these approaches generally don't take into account the semantic content of e-mails.

In this paper, we propose a novel anti-spam system which utilizes visual clues, in addition to semantic analysis information in the email body, to determine whether a message is spam.

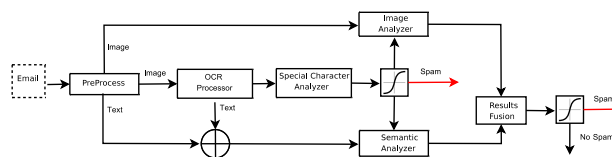


Figure 1: System Architecture

As shown in figure 1, we propose a system that integrates image and text semantic analysis using a set of hierarchical decision systems: in particular, the different modules of the algorithm are activated only when the previous phases are not sufficient to give a positive answer. We thus apply to the text contained into e-mails some semantic analysis and Natural Language Processing (NLP) based techniques, i.e.: (i) Latent Dirichlet Analysis (LDA)[7], in order to define a distribution of keywords for semantically categorizing e-mails contents; (ii) Latent

Semantic Analysis and Indexing (*LSA* and *LSI*) [3] in order to define a semantic vector space and for indexing aims. We also use Part of Speech (*POS*) analysis in order to divide the semantic problem into three simpler sub-problems, for performance reasons. Each document is represented with three sub-vectors, containing *nouns*, *verbs*, and *adjectives* respectively. In fact, it's an experimental evidence that this kind of tokens is semantically representative of the e-mails contents, obtaining all the benefits offered by *LSA*, without losing of performance. Since visual information is becoming more prevalent in emails, it becomes increasingly necessary to use such information to achieve high accuracy for anti-spam filtering. In this paper, we have analyzed a number of images contained into spam e-mails. First, we have noted that the most of spam images are artificially generated and contain embedded text (i.e. text boxes embedded into image files). We also note that Optical Character Recognition (*OCR*) tools can be efficiently used only whenever no particular content obscuring techniques are used.

For this reason, we propose to use both the visual features and other special features obtained from the *OCRs* outputs, i.e., *image features* and *OCR features*. In particular, for the first category we propose classic features such as: *Contrast*, *Entropy*, *Energy*, *Homogeneity* and *Dissimilarity*, in addition to special anti-spam features as proposed in [2]: *Perimetric Complexity*, *Noise Frequency*. About the *OCR* features, we propose the following: (i) number of special characters; (ii) maximum length of a consecutive special characters string; (iii) number of words that appear also in a dictionary; (iv) ratio between the number of special characters and normal ones; (v) ratio between the number letters in a correct words and total number of letters; (vi) ratio between the *perimetric complexity*, described before, and the total number of letters; (vii) ratio between the *perimetric complexity* and real words that are present in the text.

In order to test the efficiency of our proposal, several experiments have been carried out. First, we built a databases with 8000 spam e-mails (7000 in English), taken from the mail server of University Of Naples "Federico II". The 80% of the corpus is labeled in five main categories: Medicine, Software, Porn, Commercials, Betting Website Advertising. We also built a "legitimate" e-mail corpus consisting of 500 mail taken from wikipedia articles. We have then compared the results of our method with the ones produced by commercial anti-spam tools.

References

- [1] Ching-Tung Wu, Kwang-Ting Cheng, Qiang Zhu, Yi-Leh Wu, *Using visual features for anti-spam filtering*, Proc. IEEE Conference on Image Processing, vol. 3, pp 509-512, 2005
- [2] G. Fumera, I. Pillai, F. Roli, *Spam filtering based on the analysis of text information embedded into images*, Journal of Machine Learning Research (special issue on Machine Learning in Computer Security), 7:2699-2720, 2006
- [3] Wilfried Gansterer, Andreas Janecek, and Robert Neumayer, *Spam filtering based on latent semantic indexing*, in Proceedings of the 5th SIAM Workshop on Text Mining, Minneapolis, MN, USA, April 26-27 2007.
- [4] K.R. Gee, *Using Latent Semantic Indexing to Filter Spam*, SAC, pp 460-464, 2003.
- [5] V. Metsis, I. Androustopoulos, G. Paliouras, *Spam Filtering with Naive Bayes - Which Naive Bayes?* In Proceedings of the second Conference on Email and Anti-Spam (CEAS), Mountain View, CA, USA, 2006.
- [6] H. Drucker, with D. Wu and V. Vapnik, *Support Vector Machines for Spam Categorization* IEEE Trans. on Neural Networks, vol 10, number 5, pp. 1048-1054. 1999.
- [7] David M. Blei and Andrew Y. Ng and Michael I. Jordan, *Latent Dirichle Allocation*, Journal of Machine Learning Research, vol 3, year 2003, pages 993-1022.
- [8] Blanzieri, Enrico and Bryl, Anton *A Survey of Anti-Spam Technique* (2006) Technical Report DIT-06-056, Informatica e Telecomunicazioni, University of Trento.