Towards Symbiotic Spam E-mail Filtering

Clotilde Lopes[1] and **Paulo Cortez**[1] and **Pedro Sousa**[2]

Abstract. This position paper discusses the use of symbiotic filtering, a novel distributed data mining approach that combines content-based and collaborative filtering for spam detection.

1 INTRODUCTION

Unsolicited e-mail (spam) is a serious problem, as it consumes resources (e.g. time spent reading unwanted messages) and it is also used to spread malicious content (e.g. viruses). Currently, spam represents around 80-90% of all email messages sent [11].

Several solutions have been proposed to fight spam under two main categories [6]: Collaborative Filtering (CF) and Content-Based Filtering (CBF). CF involves sharing information about spam emails messages (e.g. blacklists with IP addresses of known spammers). The use of social networks has also been proposed to increase the CF potential (e.g. propagation of whitelists among socially connected users) [6]. CBF is the most used anti-spam solution and it is based on using a text classier (e.g. Naive Bayes) that learns to detect spam from message features (e.g. word frequencies) that are extracted from the past messages of a given email account [9].

Both pure CBF and CF solutions have drawbacks. CBF performance is poor for new users, as it requires a large number of representative examples. Also, there may be a large gap between high-level concept (e.g. spam image) and the low-level message features (e.g. bit colors). Moreover, CBF is vulnerable to contamination attacks, where spammers mix spam with normal words. On the other hand, CF suffers from first-rater, i.e. difficulty to classify emails that have not been rated before, and sparsity of data, i.e. when users rate few items. Also, spam is ultimately a personal concept and often CF systems discard this issue [7]. In recent work, we proposed the novel Distributed Data Mining (DDM) approach, called Symbiotic Filtering (SF), which combines useful features from both CBF and CF [4]. In this position paper, we discuss several SF issues.

2 SYMBIOTIC FILTERING (SF)

Symbiosis is a close interaction among different entities and this phenomenon is present not only in biological species but also in business enterprises. Under the Web 2.0 concept, the idea is to use the Internet to gather distinct users interested on similar but not identical Data Mining (DM) goals. While SF could be applied to other personalized filtering applications (e.g. web pages with offensive content), we discuss here the case of spam e-mail filtering, since it involves several challenging issues (e.g. concept drift and privacy). It is assumed that each individual runs a local CBF (e.g. Naive Bayes). Instead of sharing data (which rises privacy issues), SF exchanges information learned locally (e.g. the filter). This reduces communication costs, while preserving privacy. The final SF goal is to foster mutual relationships, where all (or most of) users benefit from the collaboration.

SF puts emphasis on accessing (indirectly) more data rather than using more complex local filters. To achieve SF, there are two interesting sharing possibilities: relevant features or DM models (i.e. filters). In this paper, we will discuss the latter approach, which can be achieved by using ensembles, where a combination function is used to aggregate distinct predictions into a single response. As spam suffers from concept drift (i.e. the learning concept evolves through time), such aggregation function should be dynamic. In [4], we proposed a hierarchical learning ensemble, where the outputs of the distinct filters are used as the inputs of another (meta-level) learner. Hence, each user has a local meta-learner that is dynamically trained (e.g. each time a new filter is received) to get a high accuracy on the user recent past data (e.g. last 100 emails). Such approach was successfully applied to spam filtering, by using a realistic mixture of real spam and ham messages. Promising results were obtained by the SF, which outperformed a local CBF under several scenarios (e.g. using fixed and incremental symbiotic groups) and for a small number of users (from 3 to 5). The increase of performance was particularly high for new users. Also, SF was more robust to word contamination attacks. By dynamically combining filters from distinct users, we believe that a stronger protection is achieved against spam. Next, we discuss in detail several important SF issues.

2.1 SF and Distributed Data Mining

The main DDM goal is to obtain a global model by aggregating several local data mining analysis [10]. SF can be viewed as a DDM variant. SF is a natural distributed response to spam, as it reuses data mining models that are already locally available at the user level. Yet, SF is distinct from classic DDM since each user will have an aggregated model that is tailored to her/his needs. Thus, there is a different motivational issue, since if a given user does not benefit from the SF, she/he could easily leave the collaboration.

2.2 SF and Other Hybrid CBF-CF Systems

SF is different from other hybrid CBF-CF works that were devised for centralized systems (e.g. collaborative ensemble learning) [13]. In SF, the data and filters are distributed through different entities, thus there are issues not only of user motivation but also distribution (how and what should be shared), privacy and security (e.g. users are not willing to share legitimate email).

2.3 Privacy and Security

Sharing DM models is less sensitive than exchanging data, but there are still privacy issues. If user A has access to the filter of B, then

[1] Dep. Information Systems, University of Minho, Portugal, email: {clopes, pcortez}@dsi.uminho.pt

[2] Dep. Informatics, University of Minho, Portugal, email: pns@di.uminho.pt

A may feed a given token (or set of tokens) into the model and thus know with some probability that such token was classified by B as spam or ham. Our privacy solution resides in an anonymous exchange of models. Privacy can be increased if each individual does not know the symbiotic group composition. Yet, for social networks it may also be attractive that the symbiotic group composition could be assessed by all the SF participants. To harvest an individual e-mail may be easy for spammers, yet knowing who belongs to a given SF group is more difficult, as security can be increased by using encrypted SF communication. Also, the use of trust weights can be used to prevent fake users from joining SF groups [12].

2.4 Communication and Scalability

When compared to CF, SF requires less communication costs. For example, a filter built from a millions of emails can be described by a few hundreds or thousands of bytes (depending on the filter algorithm used) [5]. To exchange the filters, a standard format should be adopted, such as the Predictive Model Markup Language (PMML) [8], which is compatible with a large number of data mining tools. Also, the SF exchange (e.g. filters) can be set in an asynchronous fashion and triggered only after receiving m messages or when it is detected that the filter suffers a concept shift [3].

The blind exchange of features or filters can be set using two implementations: a centralized server or a Peer-to-Peer (P2P)-like application. Under the first option, all users register into a centralized and secure service. This service could be implemented by large companies or email providers (e.g. Gmail or Hotmail), when all emails are stored at a given server. The second distributed implementation is more natural for the SF concept (e.g. under social networks), although it requires the users to trust the P2P software.

SF is linearly affected by the number of "foreign" filters. To improve scalability, we propose using text clustering [2](e.g. using the CBF word features) to segment users with similarities, thus reducing the SF group and avoiding the inclusion of inadequate or malicious filters. Another option is the use social networks, where users could choose their "friends". Such system could be implemented in social network sites (e.g. Facebook).

2.5 Current and Future Work

To experimentally test the SF capabilities, we are currently approaching the Enron corpus [1], which includes real ham messages from 148 users. Using the perl language, we defined the Enron relationships based on the number of received/sent messages between any two Enron users. Figure 1 shows an example of the social network related to user allen.p. Using a methodology similar to [4], we will select real spam from public repositories (e.g. http://untroubled.org/spam/) and use a realistic and controlled mixture of ham and spam (e.g. using different and changing spam/ham ratios). Next, we intend to apply CBF to each Enron user and test several SF approaches (e.g. text clustering and use of the Enron social network), in order to measure performance and scalability issues. At a second stage, we intend to encode SF in e-mail clients (e.g. Thunderbird extension), in order to gather feedback from real users.

3 CONCLUSION

Rather than improving a single classifier or exchanging messages, the emphasis of SF is on accessing information learned from other filters to improve personal filtering. In this position paper, we discuss several SF issues when applied to spam e-mail detection.

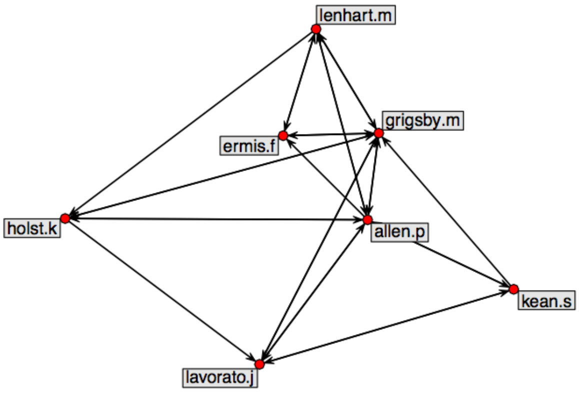

Figure 1. Example of the Enron social network

ACKNOWLEDGEMENTS

The work is supported by FCT project PTDC/EIA/64541/2006 and FCT grant SFRH/BD/44840/2008.

REFERENCES

[1] R. Beckermann, A. McCallum, and G. Huang, 'Automatic categorization of email into folders: benchmark experiments on Enron and SRI corpora', Ir-418, University of Massachusetts Amherst, (2004).

[2] R. Bilisoly, *Practical text mining with Perl*, Wiley Publishing, 2008.

[3] G. Castillo and J. Gama, 'Adaptive Bayesian network classifiers', *Intelligent Data Analysis*, **13**(1), 39–59, (2009).

[4] P. Cortez, C. Lopes, P. Sousa, M. Rocha, and M. Rio, 'Symbiotic Data Mining for Personalized Spam Filtering', in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI-09)*, pp. 149–156. IEEE, (2009).

[5] A. Garg, R. Battiti, and R Cascella, 'May I borrow your filter? Exchanging filters to combat spam in a community', in *Advanced Information Networking and Applications, 2006. AINA 2006. 20th International Conference on*, volume 2, (2006).

[6] S. Garriss, M. Kaminsky, M.J. Freedman, B. Karp, D. Mazières, and H. Yu, 'RE: reliable email', in *Proceedings of the 3rd conference on Networked Systems Design and Implementation (NSDI)*, pp. 297–310, San Jose, CA, (2006). USENIX Association Berkeley, CA, USA.

[7] A. Gray and M Haahr, 'Personalised, Collaborative Spam Filtering', in *1st Conference on E-Mail and Anti-Spam CEAS*, (2004).

[8] R. Grossman, M. Hornick, and G. Meyer, 'Data Mining Standards Initiatives', *Communications of ACM*, **45**(8), 59–61, (2002).

[9] T. Guzella and W. Caminhas, 'A review of machine learning approaches to Spam filtering', *Expert Systems with Applications*, **36**, 10206–10222, (2009).

[10] F. Provost, *Advances in Distributed and Parallel Knowledge Discovery*, chapter Distributed data mining: Scaling up and beyond, MIT Press, 2000.

[11] M86 Security Team, 'Security Labs Report', Tech. report – jul-dec 2009, available at: http://www.m86security.com/labs, Orange, CA, USA, (January 2010).

[12] H. Yu, M. Kaminsky, P. Gibbons, and A. Flaxman, 'Sybilguard: Defending against sybil attacks via social networks', *IEEE/ACM Transactions on Networking (TON)*, **16**(3), 576–589, (2008).

[13] K. Yu, A. Schwaighofer, V. Tresp, W. Ma, and H. Zhang, 'Collaborative Ensemble Learning: Combining Collaborative and Content-Based Information Filtering via Hierarchical Bayes', in *19th International Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 353–360. ACM, (2003).