Strategies and Algorithms for the Effective Control of E-mail Spam

Synopsis of the Thesis

Submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

By

K MANJUSHA

Under the Supervision of

Prof. Rahul Banerjee

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN) INDIA
2015

Introduction

Electronic mail (E-mail) has become the lifeline of the majority of modern business as well as a common vehicle for interpersonal communication between connected people. E-mail is so popular, since it is simple, cost effective and supports nearly instantaneous delivery. With such an increase in use of E-mail as a means of communication, the volume of unwanted e-mail messages (mostly spam E-mail) that is received, annually, has also grown significantly. Spam E-mails have begun to gradually undermine the integrity of E-mail and degrade online experience. With this rapid growth in E-mail spam, the financial costs as well as many other factors like involved security risks have become very significant and therefore burdensome to individuals as well as organizations. Spam emails often arrives in varying forms of continuous or near-continuous, high-volume, fast and time-varying data streams which quietly adapt to any countermeasures. Naturally, a solution has to be found to this menace of spam-mail from otherwise future of the E-mail itself may be at stake. Correct classification of such data in a dynamic environment with frequent updates, therefore, poses a major challenge.

Scope and Objectives of the Present Research

The presented work focuses on devising new strategies and algorithms that could reduce spam e-mail. The line work was chosen after exhaustive study and analysis of the contemporary e-mail spam techniques as well as relevant anti-spam mechanisms. It also involves development of efficient methods and their implementation aspects. The scope of the presented work does not include creation of a ‘ready for use’ production-grade software. Also, non-textual components of E-mails have been kept out of the scope of work.

The specific objectives of the study were:

- Devising and developing new strategies and consequent algorithms for effective control of E-mail spam; and
- Testing the resultant algorithms by the way of implementing and building a ‘proof of concept’ laboratory-scale software system in a controlled environment.

Motivation behind this Work

The primary motivation of this work is to help prevent or minimize E-mail spam so as to not only reduce the inconvenience caused to individual users by such undesired communication, but also help organizations and corporations to minimize the loss of productivity caused by such spam e-mails. The presented work, therefore, attempts to help achieve such a reduction in spam mails by the way of new and improved mechanisms for effective control of E-mail spam.

Research Gaps

Following research gaps were identified in course of literature review related to E-mail spam and associated strategies:

- Absence of any reasonably accepted single set of cohesive strategies that could take care of all major categories of Internet based E-mail spam,
- Absence of a scalable text-based mail spam classification strategy that could suitably scale by taking advantage of modern distributed computing resources,
- Near-absence of dynamic textual pattern based credible spam filters,
- Absence of computationally efficient spam classification and handling strategies that could benefit from a combination of techniques readily available in the world of computational intelligence and concurrent programming.

Research Contributions

The first algorithm 'BNNC' presented in this thesis involves a hybrid approach of Bayesian and Neural Network (NN) for classification and Genetic Algorithm for training the NN. In this case, neural network was trained with a specific Genetic Algorithm to speed up the training process. This strategy has helped since neural network training is slower, but good in terms of efficiency; whereas, genetic algorithm is good at optimization relevant to this kind of problem space. As a result, careful combination of these two mechanisms leads to both computing efficiency and less training time. This approach yields better results compared to simple classification. In this, both the header and content parts of the E-mail were considered for experimentation, which, by itself, is a complex job. Unlike statistical filters, NNs could quickly identify spam and made the classification process simpler. NNs trained with GA were used next. This helped in reducing the training and testing time. The approach has shown a high efficiency level compared with Bayesian approach. As complete information about the E-mail is taken into consideration in this approach, error rates have been found to be substantially low.

The second algorithm ‘BDT-MSVM’ presented here is a combination of Support Vector Machine (SVM) and Decision Tree. SVM proved very efficient in text classification. It provides high accuracy but is very slow in terms of training time. Decision Trees, on the other hand, classify new instances faster compared to SVMs while SVMs outperform Decision Trees in terms of accuracy. In order to exploit the advantages of both the algorithms, the strategy evolved employs a combined approach of SVM and DT to classify E-mails. This is a multi-class approach, where a new class called ‘likelihood (of) spam’ is identified. Normal data points were classified with the help of decision tree, whereas some crucial data points were classified by employing multi-class SVM. When compared with other classifying algorithms/method this approach demonstrated reasonably higher performance in terms of accuracy without compromising in terms of the time complexity.

The third algorithm ‘ES^{2}C’ presented here uses an ensemble approach of Map-Reduce based SVM. The intention of this ensemble approach was to reduce training time of the filtering process as well as maintain the high degree of accuracy. SVM was employed as a classifier, as it offers high classification accuracy. This alone, however, was not a good choice since it needs more training time for high dimensional data. With the Map-Reduce process, the training time can be decreased dramatically and this way in this algorithm SVM and Map-Reduce were strategically combined for use.

The fourth and the last algorithm ‘SpamReduce’ presented in this thesis leads to a parallel and distributed scheme based classification. K-Nearest Neighbourhood (kNN) Join algorithm was used for classification since it works well for multi-dimensional data. A parallel distributed environment was created to reduce the dimensionality of the data. This approach was devised with a distributive strategy that exploits parallel characteristics of kNN which further improved the resultant classification efficiency.

Both of these algorithms were based on Map-Reduce for large scale E-mail classification using ensemble based SVM and kNN join, since it has a significant potential for parallelism. The most attractive feature of these strategies lies in its simplicity and flexibility that enable it to be implemented over any parallel/distributed paradigm. These strategies were able to deliver promising results that make these suitable for large scale E-mail classification for significant improvement in E-mail spam reduction.

Organization of the Thesis

The content of this thesis has been organized in the form of five chapters.

Chapter 1: Introduction

This is an introduction to the research area (E-mail spam) and provides an overview of the principal motivation, objective and scope of the work done.

Chapter 2: Literature Review

This chapter presents a thorough literature survey which involves due analysis of anti-spam strategies, examining the types of solutions available so far for controlling E-mail spam as well as related algorithms. This chapter also analyses resource requirements of existing algorithms vis-à-vis ways in which they carry out identification of the reasons. It also lists select situations which may render the prevalent strategies and algorithms less effective or ineffective.

Chapter 3: Modeling the E-mail Spam

In this chapter, brief experimental modeling of E-mail spam has been carried out along with discussions providing the basis involved and any alternative approach against which the chosen method stands out.

Chapter 4: Strategies and Algorithms

This chapter discusses the chosen strategies, schemes and methods for controlling spam and analysis of their effectiveness and presents corresponding algorithms along with evidence of their performance vis-à-vis known algorithmic approaches and strategies.

Chapter 5: Conclusions

This chapter presents the conclusions drawn on the results obtained from the experiments, compares them briefly with other related algorithms and also summarizes both, the principal contributions of the work done as well as its limitations. Planned future work has also been briefly presented at the end.

Limitations of the Work Done

The principal limitation of the work done is that in view of the scope of the work, the work only targets the textual elements of E-mails and does not consider attached media in multiple formats. Also, scalability of the presented solution has been provisioned but has only been verified on a local scale using multiple computing nodes spread over multiple physical systems and not in a live e-mail environment. However, by involving three public and one private spam mail data-sets, it was possible to state with confidence that the strategies and algorithms presented here would prove useful in live environments as well.

Future Work

Future work for this study might involve different aspects, like focus towards increasing the efficiency of the identified and developed filters; improving implementation (of the strategies) and exploiting more potential parallelism. Support Vector Machine for spam classification also deserves more attention. The work done in SVM was only to a limited extent, but can be further explored by using the kernel, among other things, for classification. Another possible idea is to extend this work to consider other types of spam like image spam and PDF spam. In addition, issues related to attachments may also be addressed in the subsequent stages, for the sake of overall utility in real-life conditions, even though both of these elements have nothing to do with textual classification.

List of Select References

Alham, N. K., Li, M., Liu, Y., & Qi, M. A MapReduce-based distributed SVM ensemble for scalable image classification and annotation. Computers & Mathematics with Applications, vol.66, no.10, pp:1920-1934. 2013.

Barracuda Reputation Block List (BRBL); Available online at: http://www.barracudacentral.org/rbl Last accessed: December 1, 2015.

Böhm, C. and Krebs, F. The k-nearest neighbour join: Turbo charging the KDD process. Knowledge and Information Systems, vol.6 no.6, pp.728-749. 2000.

Caruana, Godwin, Maozhen Li, and Man Qi. A MapReduce based parallel SVM for large scale spam filtering. In Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on, vol. 4, pp. 2659-2662. 2011.

Cortes, Corinna, and Vladimir Vapnik. Support-vector networks. Machine learning vol.20, no. 3 pp: 273-297. 1995.

Drucker, Harris, Donghui Wu, and Vladimir N. Vapnik. Support vector machines for spam categorization. Neural Networks, IEEE Transactions on 10.5 pp: 1048-1054. 1999.

González-Talaván, Guillermo. A simple, configurable SMTP anti-spam filter: Greylists. computers & security vol.25, no. 3 pp: 229-236.2006.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter Vol.11, no. 1 pp: 10-18. 2009.

Hsu, Chih-Wei, and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. Neural Networks, IEEE Transactions on vol.13, no. 2 pp: 415-425. 2002

Klimt, Bryan, and Yiming Yang. The Enron Corpus: A new dataset for email classification research. In Machine learning: ECML 2004, Springer, Berlin Heidelberg, pp. 217-226. 2004

Lewis, David D., and Marc Ringuette. A comparison of two learning algorithms for text categorization. In Third annual symposium on document analysis and information retrieval, vol. 33, pp. 81-93. 1994.

Manjusha, K; Shahabas, S; Banerjee, Rahul: An Efficient Method of Spam Classification by Multi-class Support Vector Machine Classifier, Proceedings of the Seventh International Conference on Data Mining and Warehousing ICDMW, pp: 102-110, 2013.

Meyer, Oliver, Bernd Bischl, and Claus Weihs. Support vector machines on large data sets: Simple parallel approaches. In Data Analysis, Machine Learning and Knowledge Discovery, Springer International Publishing, pp. 87-95. 2014.

Sahami, M, Dumais, S, Heckerman, D and Horvitz, E., A Bayesian Approach to Filtering Junk E-Mail , AAAI-98 Workshop on Learning for Text Categorization, Vol. 62, pp. 98-105 1998.

Wang, J., Gao, K., & Vu, H. Q. SpamCooling: a parallel heterogeneous ensemble spam filtering system based on active learning techniques. Journal of convergence information technology, vol. 5, no. 4, pp: 90-102. 2010.

Xia, Chenyi, Hongjun Lu, Beng Chin Ooi, and Jing Hu. Gorder: an efficient method for KNN join processing. In Proceedings of the Thirtieth international conference on Very large data bases- Volume 30, pp. 756-767. 2004.

Yang, Zhen, Xiangfei Nie, Weiran Xu, and Jun Guo. An approach to spam detection by naive Bayes ensemble based on decision induction. In Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on, vol. 2, pp. 861-866. 2006.

Ying, K. C., Lin, S. W., Lee, Z. J., & Lin, Y. T. An ensemble approach applied to classify spam e-mails. Expert Systems with Applications, vol. 37, no. 3, pp. 2197-2201. 2010.

Youn, Seongwook, and Dennis McLeod. A comparative study for email classification. In Advances and Innovations in Systems, computing sciences and software engineering, Springer Netherlands, pp. 387-391. 2007.

Zhao, Jun, Zhu Liang, and Yong Yang. Parallelized incremental support vector machines based on MapReduce and Bagging technique. In International Conference on Information Science and Technology (ICIST), IEEE pp: 297-301. 2012.

Zhang, Le, Jingbo Zhu, and Tianshun Yao. An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing (TALIP) vol. 3, no. 4 pp : 243-269. 2004.