© 2018 IJSRSET | Volume 4 | Issue 11 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET

A Review Various Techniques for Content Based Spam Filtering

Minhaz Fatima Nayanmulla Kallu Pathan, Prof. Vijaya Kamble
M. Tech CSE, Guru Nanak Institute of Engineering & Technology, Nagpur, Maharashtra, India

ABSTRACT

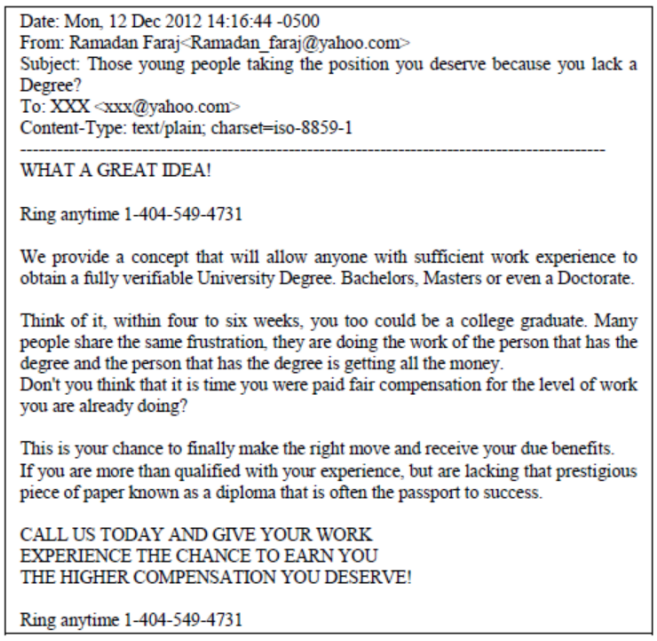
In recent years' spam became a major problem of Internet and electronic correspondence. There developed plenty of techniques to battle them. In this paper, the overview of existing e-mail spam filtering methods is given. The classification, evaluation, and correlation of conventional and learning-based methods are provided. Some personal enemy of spam items is tested and compared. The statement for a new methodology in spam filtering technique is considered.

Keywords : Spam Filtering, Machine learning, Learning-Based Methods, Classification

I. INTRODUCTION

In recent years, e-mails have become a typical and critical medium of correspondence for most Internet users. However, spam, otherwise called unsolicited commercial/mass e-mail, is a bane of e-mail correspondence. Spam is generally compared to paper junk mail. However, the difference is that junk mailers pay a fee to distribute their materials, whereas with spam the recipient or ISP pay as extra data transfer capacity, plate space, server resources, and lost profitability. In the event that spam continues to develop at the current rate, the spam problem may become unmanageable in the near future.

An examination estimated that over 70% of the present business emails are spam [1]; therefore, there are numerous serious problems associated with developing volumes of spam, for example, filling users' mailboxes, engulfing critical personal mail, squandering storage space and correspondence data transmission, and expanding users' time to delete all spam emails. Spam emails shift altogether in content and they generally belong to the accompanying categories: money making tricks, fat misfortune, improve business, sexually explicit, make friends, service provider advertisement, etc.[2], One example of a spam mail appears as Fig. 1.



Date: Mon, 12 Dec 2012 14:16:44 -0500
From: Ramadan Faraj<Ramadan_faraj@yahoo.com>
Subject: Those young people taking the position you deserve because you lack a Degree?
To: XXX <xxx@yahoo.com>
Content-Type: text/plain; charset=iso-8859-1

WHAT A GREAT IDEA!

Ring anytime 1-404-549-4731

We provide a concept that will allow anyone with sufficient work experience to obtain a fully verifiable University Degree. Bachelors, Masters or even a Doctorate.

Think of it, within four to six weeks, you too could be a college graduate. Many people share the same frustration, they are doing the work of the person that has the degree and the person that has the degree is getting all the money.
Don't you think that it is time you were paid fair compensation for the level of work you are already doing?

This is your chance to finally make the right move and receive your due benefits. If you are more than qualified with your experience, but are lacking that prestigious piece of paper known as a diploma that is often the passport to success.

CALL US TODAY AND GIVE YOUR WORK
EXPERIENCE THE CHANCE TO EARN YOU
THE HIGHER COMPENSATION YOU DESERVE!

Ring anytime 1-404-549-4731

Figure 1. An Example of Spam E-mail

E-mail users spend an increasing measure of time reading message and deciding whether they are spam or not and categorizing them into folders. E-mail service providers might want to relieve users from this burden by introducing server-based spam filters

that can group e-mails as spam consequently. [3] Spam filtering classification due to the accompanying reasons:

- Continually changing – Spam is always showing signs of change as spam on new themes emerges. Likewise, spammers attempt to make their messages as indistinguishable from legitimate email as could be expected under the circumstances and change the patterns of spam to thwart the filters. [4]
- False positives problem – false positives are just unacceptable; along these lines, the requirements on the spam filter are very exacting.
- OCR computational expense – the OCR computational expense in-text embedded in images compatible with the huge measure of e-mails handled every day by the server-side filter. [4]
- The use of content darkening techniques – Spammers are applying content clouding

II. REVIEW OF SPAM FILTERING METHODS

In spite of the fact that the primary spam was sent in 1978, it began to be written about it as a problem in scientific literature just from 1982. One of the principal papers where this problem is considered is Peter J. Denning's article [4]. The principal mathematical device applied to spam filtering systems is the Bayes' calculation, which was used first by Sahami et.al in 1996 and after that by other researchers [5-8]. Bayes' classifier relies on well-known Bayes theorem and the primary papers about it could be met as early as 1960 [9]. Amid more than 40-year history, Naive Bayes Classifier (NBC) was used for the arrangement of very different type of undertakings: from the classification of texts in news agencies until essential finding of diseases in medicine. For the problems where NBC is applied, there is normally selected presence or absence of words in the text as a characteristic, i.e. the set of characteristics T is a set off all words in documents. Hereby, if the word t_i is present, the weight of characteristics $w_i=1$, otherwise wi=0. In the case of e-mail filters where spam classification is used, there taken into the record the area where the word had been met: heading, subject, and body of the e-mail.

Beginning from the distribution of Gary Robinson [10], in some filters (for example, Spam Assassin) there came to be used the method of overlapping probabilities suggested by R. Fisher in 1950. For spam detection, Robin-child offered to calculate not just the likelihood of "spamness" of the document, yet in addition the likelihood of "legitimness" of email. The next directions were the application of Markov chain PageRank and Hidden Markov Model which are met in papers Paolo B., et al. [11], and José Gordillo, et al. [12]. Kolmogorov complexity estimation is met in papers Spracklin L.M., et al. [13]. Stomach muscle absolutely another methodology is a new method of advanced examination of textual e-mails for spam detection which can be right off the bat observed in paper Korelov S. V., et al. [14]. Here e-mail is considered as a flag x(n), after the methods of computerized processing are applied to signals and the likelihood of false positives are defined for these methods. Utilization of methods of clustering analyses to the problem of filtering e-mails to legitimate and spam is considered in papers [15-18]. From the 2009 year, beginning from Paulo Cortez's, et al. article [19] one can meet the statement as a Symbiotic Data Mining which is a cross breed of Collaborative Filtering (CF) and Content-Based Filtering (CBF).

Considering shocking measure of spam messages coming to e-mail boxes it is possible to assume that spammers operate not alone, there are worldwide, organized, virtual informal organizations of spammers. They assault e-mails of not just users, even whole partnerships and countries. Spam is of the weapons of data war. In spite of the way that, the terms spam and war appear in one context [20,21] since the 2003 year, just from 2009, the problem of spammers' informal organizations are considered in scientific papers.

Clustering of spammers considering them in gatherings is offered in paper Fulu Li, et al. [22]. In works Xu K.S., et al. [23,24] the method of spectral clustering is applied to the set of spam messages collected under project Honey Pot for defining and following of interpersonal organizations of spammers.

They represent an interpersonal organization of spammers as a diagram the gestures of which correspond to spammers, and a corner between two intersections of the chart as social relations between spammers.

Research and development of spam filtering systems are actively carried everywhere throughout the world. Alongside scientific institutes, there are numerous associations and corporations investigating and offering different theoretical, commonsense and juridical approaches to spam filtering. Different associations as university laboratories (laboratories CSAIL MIT in USA [25], Computer Laboratory Faculty University of Cambridge in UK [26] and etc.); research centers (NCSR Democritos in Greece [27], research center of IBM [28,29] and etc.); commercial companies (Microsoft [30], Symantec [31], Kaspersky's Laboratory [32] and etc.) had been involved to this process. Numerous international associations take great attention to the concerned problem. It is created the ASRG (Anti-Spam Research Group) [33] inside the association IETF (Internet Engineering Task Force) [34] in 2003.

III.CONCLUSION

After the study of above-listed literature, we come to the following conclusion. Spammers constantly change external signs of e-mails to skip spam filtering systems, there arises a need for adaptive filtering system, which should have the ability to react quickly to the changes and provide fast and qualitative self-tuning in accordance with a new set of features.

Since the filters are trained on a very limited number of messages that come only to a specific user or a specific mail provider, the quality of filtration in the existing client and server filtering systems is rather low. But it can be improved if to apply the hybrid filtration system, in other words, the complex hierarchical and multi-agent filtration system that helps users to participate in the identification of the filtering errors and the appropriate setting of filters at each level (user level, organization level, mail provider level).

Therefore it is quite perspective for solving this problem, the combination of two widespread approaches as using the personal e-mail classification model on a server-side solution. Development of server-side personalized e-mail filtering systems that use the learning-based classification algorithms based on Data Mining methods is a very perspective direction.

This statement is supported by the followings:

- Personalized server-side filtering systems are preferable than the client side solutions because provide universal access to an e-mail, reduce expenses, which is very important for corporate users;
- Personalized server-side filtering systems are more preferable because of greater accuracy and fewer errors in comparison with the general model;
- Personalized server-side filtering system offered in the author's another paper based on the Universal Declaration of Human Rights and has a universal character, can be applied in all countries; learning-based algorithms used in personalized server-side filtering systems exceed traditional ones because of a number of fundamental qualities (quality of filtering, the absence of updates, autonomy, independence from external knowledge bases).

IV.REFERENCES

[1] Aladdin Knowledge Systems, Anti-spam white paper,Retrieved December 28, 2011.

[2] F. Smadja, H. Tumblin, "Automatic spam detection as a text classification task", in: Proc. of Workshop on Operational Text Classification Systems, 2002.

[3] A. Hassanien, H. Al-Qaheri, "Machine Learning in Spam Management", IEEE TRANS., VOL. X, NO. X, FEB.2009

[4] P. Cunningham, N. Nowlan, "A Case-Based Approach to Spam Filtering that Can Track Concept Drift", Retrieved December 28, 2011

[5] M. Sahami, "Learning Limited Dependence Bayesian Classifiers," Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, The AAAI Press, Menlo Park, 1996, pp. 334-338.

[6] M. Sahami, S. Dumais, D. Heckerman and E. Horvitz, "A Bayesian Approach to Filtering Junk Email," AAAI Technical Report WS-98-05, AAAI Workshop on Learn-ing for Text Categorization, 1998.

[7] J. R. Hall, "How to Avoid Unwanted Email," Communi-cations of the ACM, Vol. 41, No. 3, 1998, pp. 88-95. doi:10.1145/272287.272329

[8] E. Gabber, M. Jakobsson, Y. Matias and A.J. Mayer, "Curbing Junk E-Mail via Secure Classification," Pro-ceedings of the Second International Conference on Fi-nancial Cryptography, Springer-Verlag London, 23-25 March 1998, pp. 198-213.

[9] R. A. Fisher, "On Some Extensions of Bayesian Inference Proposed by Mr. Lindley," Journal of the Royal Statisti-cal Society: Series B, Vol. 22, No. 2, 1960, pp. 299-301.

[10] G. Robinson, "A Statistical Approach to the Spam Prob-lem," 2003. http://www.linuxjournal.com/article.php?sid=6467 (ac-cessed March 2011).

[11] P. Boldi, M. Santini and S. Vigna, "PageRank as a Func-tion of the Damping Factor," Proceedings of the 14th In-ternational Conference on World Wide Web, ACM New York, 10-14 May 2005. doi:10.1145/1060745.1060827

[12] J. Gordillo and E. Conde, "An HMM for Detecting Spam Mail," Expert Systems with Applications, Vol. 33, No. 3, 2007, pp. 667-682. doi:10.1016/j.eswa.2006.06.016

[13] L. M. Spracklin and L. V. Saxton, "Filtering Spam Using Kolmogorov Complexity Estimates," in Russian, 21st In-ternational Conference on Advanced Information Net-working and Applications Workshops (Ainaw'07), Niag-ara Falls, 21-23 May 2007, pp. 321-328.

[14] S. V. Korelov, A. K. Kryukov and L. U. Rotkov, "Text Messages' Digital Analysis on Spam Identification," in Russian, Proceedings of Scientific Conference on Radio-physics, Nizhni Novgorod State University, Nizhny Nov- gorod Oblast, 2006.

[15] W.-F. Hsiao and T.-M. Chang, "An Incremental Clus-ter-Based Approach to Spam Filtering," Expert Systems with Applications, No. 34, No. 3, 2008, pp. 1599-1608. doi:10.1016/j.eswa.2007.01.018

[16] S. M. Lee, D. S. Kim and J. S. Park, "Spam Detection Using Feature Selection and Parameters Optimization," IEEE International Conference on Intelligent and Soft-ware Intensive Systems, Krakow, 15-18 February 2010, pp. 883-888. doi:10.1109/CISIS.2010.116

[17] M. F. Saeddian and H. Beigy, "Spam Detection Using Dynamic Weighted Voting Based on Clustering," Pro-ceedings of the 2008 Second International Symposium on Intelligent Information Technology Application, Vol. 2, pp. 122-126. doi:10.1109/IITA.2008.140

[18] M. Sasaki and H. Shinnou, "Spam Detection Using Text Clustering," IEEE Proceedings of the 2005 International Conference on Cyberwords, Singapore, 23-25 November 2005, pp. 316-319. doi:10.1109/CW.2005.83

[19] P. Cortez, C. Lopes, P. Sousa, M. Rocha and M. Rio, "Symbiotic Data Mining for Personalized Spam Filter-ing," IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Milan, 15-18 September 2009, pp. 149-156. doi:10.1109/WI-IAT.2009.30

[20] W. Lauren, "Spam Wars," Communications of the ACM —Program Compaction, Vol. 46, No. 8, 2003, p. 136.

[21] G. Pawel and M. Jacek, "Fighting the Spam Wars: A Re-Mailer Approach with Restrictive Aliasing," ACM Transactions on Internet Technology (TOIT), Vol. 4, No. 1, 2004, pp. 1-30.

[22] F. Li, H. Mo-Han and G. Pawel, "The Community Be-havior of Spammers" 2011. http://web.media.mit.edu/~fulu/ClusteringSpammers.pdf.

[23] K. S. Xu, M. Kliger, Y. Chen, P. J. Woolf and A. O. Hero, "Revealing Social Networks of Spammers through Spec-tral Clustering," IEEE International Conference on Com-munications, Dresden, 14-18 June 2009, pp. 1-6. doi:10.1109/ICC.2009.5199418

[24] K. S. Xu, M. Kliger and A. O. Hero, "Tracking Commu-nities of Spammers by Evolutionary Clustering," 2011.

[25] Laboratory CSAIL MIT in USA, 2011. http://projects.csail.mit.edu/spamconf/.

[26] Computer Laboratory Faculty Cambridge University in UK, 2011. http://www.cl.cam.ac.uk/~rnc1/.

[27] National Center for Scientific Research, "Demokritos," 2011. http://www.iit.demokritos.gr/.

[28] D. Mertz, "Spam Filtering Techniques," 2002. http://www.ibm.com/developerworks/linux/library/l-spamf.html.

[29] R. Segal, J. Crawford, J. Kephart and B. Leib, "Spam-Guru: An Enterprise Anti-Spam Filtering System," IBM Thomas J. Watson Research Center. http://www.research.ibm.com/people/r/rsegal/papers/spamguru-overview.pdf.

[30] Microsoft Antispam Technologies. http://www.microsoft.com/mscorp/safety/technologies/antispam/default.mspx.

[31] Symantec Antispam Protection for E-Mail. http://www.symantec.com/business/premium-antispam,.

[32] Kasperskiy Ant-Spam. http://www.kaspersky.ru/anti-spam.

[33] Anti-Spam Research Group. http://asrg.sp.am/.

[34] The Internet Engineering Task Force. http://www.ietf.org

Cite this article as :

Minhaz Fatima Nayanmulla Kallu Pathan, Prof. Vijaya Kamble, "A Review Various Techniques for Content Based Spam Filtering", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), ISSN : 2456-3307, Volume 4 Issue 11, pp. 267-271, November-December 2018.
Available at doi :
https://doi.org/10.32628/18410IJSRSET
Journal URL : http://ijsrset.com/IJSRSET21841138