IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 04-07

A Survey on Content Based Spam Filtering

Mr. Mukesh Baburao Salam, Prof. Prof. Jayant Adhikari
M-Tech student Department of Computer Science and Engineering,
Tulsiramji Gaikwad-Patil College of Engineering & Technology, Nagpur, Maharashtra, India.
ajyoti0326@gmail.com
Guide,Assistant Professor Department of Computer Science and Technology
Tulsiramji Gaikwad-Patil College of Engineering & Technology, Nagpur, Maharashtra, India.

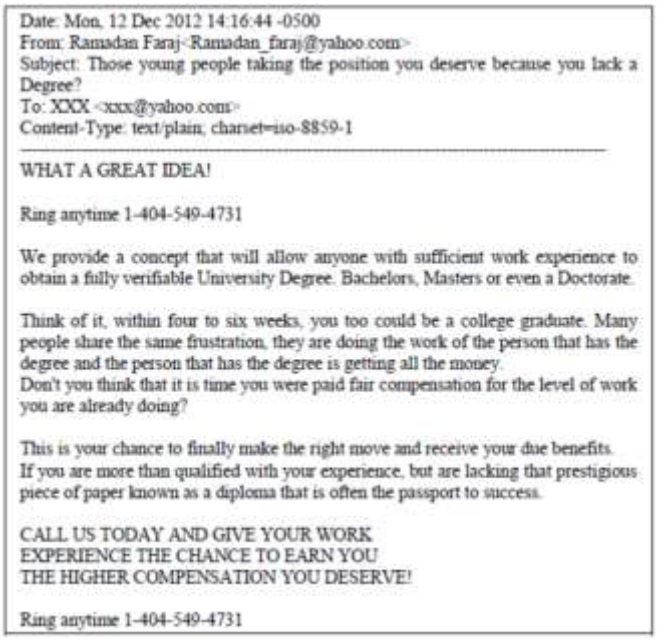
Abstract: *Email spam or junk e-mail (unwanted e-mail "usually of a commercial nature sent out in bulk") is one of the major problems of the today's Internet, bringing financial damage to companies and annoying individual users. Among the approaches developed to stop spam, filtering is an important and popular one. Common uses for mail filters include organizing incoming email and removal of spam and computer viruses. A less common use is to inspect outgoing email at some companies to ensure that employees comply with appropriate laws. Users might also employ a mail filter to prioritize messages, and to sort them into folders based on subject matter or other criteria. Mail filters can be installed by the user, either as separate programs, or as part of their email program (email client). In email programs, users can make personal, "manual" filters that then automatically filter mail according to the chosen criteria. In this paper, we present a survey of the performance of five commonly used machine learning methods in spam filtering. Most email programs now also have an automatic spam filtering function.*

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

I. Introduction

Lately, messages have turned into an average and basic mode of correspondence for most Internet clients. In any case, spam, generally called spontaneous business/mass email, is a worst thing about email correspondence. Spam is for the most part contrasted with paper garbage mail. Notwithstanding, the thing that matters is that garbage mailers pay an expense to appropriate their materials, though with spam the beneficiary or ISP pay as additional information exchange limit, plate space, server assets, and lost gainfulness. In the occasion that spam keeps on creating at the present rate, the spam issue may wind up unmanageable sooner rather than later.

An examination assessed that over 70% of the present business messages are spam [1]; in this manner, there are various significant issues related with creating volumes of spam, for instance, filling clients' letter drops, immersing basic individual mail, wasting extra room and correspondence information transmission, and extending clients' an ideal opportunity to erase all spam messages. Spam messages move by and large in substance and they by and large have a place with the going with classes: cash making traps, fat disaster, improve business, explicitly unequivocal, make companions, specialist co-op commercial, etc.[2], One case of a spam mail shows up 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. Literature Review

Notwithstanding the way that the essential spam was sent in 1978, it started to be expounded on it as an issue in logical writing just from 1982. One of the main papers where this issue is considered is Peter J. Denning's article [4]. The chief numerical gadget connected to spam sifting frameworks is the Bayes' computation, which was utilized first by Sahami et.al in 1996 and after that by different analysts [5-8]. Bayes' classifier depends on surely understood Bayes hypothesis and the essential papers about it could be met as right on time as 1960 [9]. In the midst of over 40-year history, Naive Bayes Classifier (NBC) was utilized for the course of action of altogether different kind of endeavors: from the characterization of writings in news offices until fundamental finding of infections in drug. For the issues where NBC is connected, there is ordinarily chosen nearness or nonattendance of words in the content as a trademark, for example the arrangement of attributes T is a set off all words in archives. Therefore, if the word ti is available, the heaviness of attributes wi =1, generally wi=0. On account of email channels where spam arrangement is utilized, there taken into the record the zone where the word had been met: heading, subject, and body of the email.

Starting from the dissemination of Gary Robinson [10], in certain channels (for instance, Spam Assassin) there came to be utilized the strategy for covering probabilities proposed by R. Fisher in 1950. For spam identification, Robin-kid offered to figure not simply the probability of "spamness" of the archive, yet furthermore the probability of "legitimness" of email. The following headings 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 multifaceted nature estimation is met in papers Spracklin L.M., et al. [13]. Stomach muscle totally another system is another technique for cutting edge examination of literary messages for spam identification which can be directly off the bat seen in paper Korelov S. V., et al. [14]. Here email is considered as a banner x(n), after the techniques for modernized handling are connected to signals and the probability of false positives are characterized for these strategies. Usage of techniques for bunching examinations to the issue

of separating messages to genuine and spam is considered in papers [15-18]. From the multi year, starting from Paulo Cortez's, et al. article [19] one can meet the announcement as a Symbiotic Data Mining which is a cross type of Collaborative Filtering (CF) and Content-Based Filtering (CBF).

Considering stunning proportion of spam messages coming to email boxes it is conceivable to accept that spammers work not the only one, there are around the world, composed, virtual casual associations of spammers. They strike messages of not simply clients, even entire associations and nations. Spam is of the weapons of information war. Regardless of the way that, the terms spam and war show up in one setting [20,21] since the multi year, just from 2009, the issue of spammers' casual associations are considered in logical papers. Bunching of spammers considering them in social occasions is offered in paper Fulu Li, et al. [22]. In works Xu K.S., et al. [23,24] the strategy for phantom bunching is connected to the arrangement of spam messages gathered under venture Honey Pot for characterizing and following of relational associations of spammers.
They speak to a relational association of spammers as a graph the motions of which compare to spammers, and a corner between two crossing points of the diagram as social relations between spammers.

Innovative work of spam separating frameworks are effectively conveyed wherever all through the world. Close by logical foundations, there are various affiliations and partnerships researching and offering distinctive hypothetical, presence of mind and juridical ways to deal with spam sifting. Diverse relationship as college (labs CSAIL MIT in USA [25], Computer Laboratory Faculty University of Cambridge in UK [26] and so forth.); examine focuses (NCSR Democritos in Greece [27], inquire about focal point of IBM [28, 29] and so on.); business organizations (Microsoft [30], Symantec [31], Kaspersky's Laboratory [32] and so on.) had been included to this procedure. Various global affiliations take incredible regard for the concerned issue. It is made the ASRG (Anti-Spam Research Group) [33] inside the affiliation IETF (Internet Engineering Task Force) [34] in 2003.

III. Conclusion

After the investigation of above-recorded writing, we arrive at the accompanying resolution. Spammers always show signs of change outer indications of messages to skip spam separating frameworks, there emerges a requirement for versatile sifting framework, which ought to be able to respond rapidly to the progressions and give quick and subjective self-tuning as per another arrangement of highlights.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).

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/.