Contextual Strategies for Detecting Spam in Academic Portals

Balaji Rajendran and Anoop Kumar Pandey

Centre for Development of Advanced Computing,
Electronics City, Bangalore-560100
{balaji,anoop}@cdac.in

Abstract. The emergence of social networking platforms in online space and its ever increasing user base has opened up a new arena for the spammers to exploit. Spam, in these kinds of platforms and such other interactive tools like forums, instant messaging, could be created easily and difficult to stop it from spreading, which necessitates the development of better detection strategies. In this paper, we present a contextual strategy for detecting spam in a restricted domain such as an academic portal. The proposed method uses the relationship between the concepts of the domain and the concepts of the individual message fragments to determine the relevancy of the message to the given context and marks the outliers. The strategy has been tested using a prototype system which had networking and interactive features for the participants to share information, and the results indicated that the contextual strategy was fairly successful in detecting spam.

Keywords: Contextual Strategies, Spam detection, Academic Portal, Socio-contextual, Information Systems.

1 Introduction

Spamming is the act of spreading unsolicited, unrelated and irrelevant content in the online world through various utilities such as Email, Discussion forums, Instant messaging, Social Networking and through interactive information sharing web applications. The most recognized version of spam is email spam [1-4] and from that perspective, spam can be classified as Spam without attachment and with attachment [5]. Spamming remains economically viable because advertisers have no operating costs beyond the management of their mailing lists, and it is difficult to hold senders accountable for their mass mailings.

Spam in blogs, also called simply as blog spam or comment spam is a form of spamdexing, that occurs when unrelated comments to a piece of information is posted, typically those unrelated ads found in blogs, wikis, guestbooks, and other publicly accessible online forums. Spammers in the above utilities exploit by searching for specific widgets or controls, which accept a user's information and display them, and add links to their sites of interest. This would lead to increased ranking for those sites, often misleading users and customers [9].

N. Meghanathan et al. (Eds.): CCSIT 2012, Part II, LNICST 85, pp. 250–256, 2012.

The main problems with spamming are [10]:

- Spammer campaigns result in undeserved higher ranking for dubious pages in search engine results
- Waste the time and effort of real users, by cluttering the information space
- Trick the users and damage the reputation of good systems.

Other negative impacts include involving overwhelming moderators and administrators, to find, obstruct and remove the misleading spam, in order to protect the genuine and legitimate use of systems [11-13]. Previous studies show that comment spam in online discussion forums is prevalent and techniques to counter such type of spam have attracted several researchers' attention [14-18]. Several content-based methods have been proposed to automatically identify spam comments. Content-based methods analyze the text of the post or message (such as checking for the presence of predefined terms or links) in a forum and infer the likelihood of a message being spam or legitimate.

Recently, spammers are also targeting users of social networking services such as Facebook, Orkut etc… Spammers utilize the resource sharing features in the above sites to their advantage, by embedding links to their sites of interest, often pornographic or to sites that sell something. They are also easily able to target a certain demographic segment of users, by exploiting the group or fan page facilities provided by the above sites. Though the above sites may feature a "Repot Spam" or "Report Abuse" facility, the spammers get around it by frequently changing their addresses from one account to another [19].

With the popularity of social networking sites ever increasing, the use of the similar concepts for professional networking [6], and Academic networking [7-8] have become popular. Whatever be the ways of information sharing in whatever domain, the presence of spam is prevalent. However as the domain shrinks or redefines itself within a limited domain, the spam detection methods needs to be redefined for accuracy and for more effectiveness. In this paper we are targeting the practice of spam in academic networking environment, where the prime stakeholders are academicians.

The rest of this paper is organized as follows. Section 2 discusses some of the spam detection techniques primarily in Web 2.0 environments. Section 3 details the Contextual strategies for detecting spam in academic portals. Section 4 discusses the impact of our proposed algorithm and results and section 5 concludes the paper.

2 Related Work

Heymann et al. presented a survey of approaches for fighting spam on social networking portals [12]. Hayati presented an evaluation and analysis of Web 2.0 anti-spam methods [11]. Benevenuto et al. provided a general overview of pollution in video sharing systems such as YouTube [13] with evidence of pollution, types of pollution, effect on the system and control strategies.

Research in blog spam is relatively in its infancy. One of the first articles to talk about blog spam was presented in early 2004 [20] which was limited to existence of

spam in blog. In [21] the authors proposed a collaboration spam detection method for detecting link spam inside comments and track back. Authors in [22] proposed an idea to detect blog spam based on vocabulary inside blog post, comment and track back. Methods presented in [23] involve use of supervised machine learning approach to detect spam in Blogs.

A spam detection method was presented in [24] which, employs 40 features to differentiate spam from legitimate profiles in social networking websites. It uses Naïve Bayesian machine learning algorithm to do supervised spam detection task and depend on features that can/cannot be language independent. There is no pressure on user side for differentiation among genuine users and spammers. In [25] authors proposed spam detection method for combating spam in video-sharing websites. Their supervised approach use videos' meta-data information to do the classification task. There is no increase in complexity of user-and-system interaction.

The authors in [26] proposed an idea of tagging system which, can be robust for detection of spam as it counts number of coincident (or common) tags amongst other users and assigns document a relevance ranking number. By looking at the ranking number, one can differentiate among spam and legitimate content. This method is language independent and content based. This domain of spam battle is young and hence the research in strategies is still in its nascent stages. The emergence of Web 2.0 has necessitated the development of sophisticated and unsupervised methods for spam detection.

In [27] Ashish Surekha has developed a heuristics and a solution framework with some key components like ATDC (Average Time Difference between Comments), PCHF (Percentage of Comments with hasSpamHint Flag), CRAV (Comment Repeatability Across Videos), CRR (Comment Repetition and Redundancy) for detecting potential spammers in YouTube.

Our work proposes to develop and use contextual strategies for detecting spam, as explained in the next few sections. These contextual strategies have been used to discover similar knowledge gathering tasks undertaken by users in a Web Information system [28], and also to mine such tasks for providing user assistance [29].

3 Contextual Strategies for Spam Detection

In vertical or domain-specific portals, the main stakeholders and the kind of content that could be found are well known in advance. The restricted audience for these kinds of portals is not going to stop the spammers from their attacks. However the spam detection technique could be improved with the additional knowledge of the subject domain and the users to be more precise, effective and accurate.

In this paper, we consider the special case of academic networking portals, where the main users are academicians viz. students, faculties, researchers etc, whose sole aim is to share knowledge and information about their subjects of interest, information about institutions, courses, events, projects, questionnaires, and such related activities for learning and research. We propose a spam detection technique based on contextual strategies that could be highly effective for academic domains and start by defining the entities involved in it.

3.1 Definitions

Resource: All components of academic networking websites can be treated as resources. Examples may include institutes, faculties, courses, events, projects, web links etc.

Message: Message in this context has been used in a broader sense encompassing all types of information that users share among themselves. Messages could be personal messages addressed to a particular user or could be notifications about a particular event/conference or could be questions asked on a specific topic or could be comments on an academic article or web resource.

Concepts & Relationships: Concept in general, could refer to all the terminologies and vocabulary of a particular domain which is used to describe it. The definition of concepts and relationship between the concepts is typically captured in the form of ontology for a domain. Here for the purpose of the detecting spam in academic domain, we construct a concept tree that captures only two kinds of relationships: is-a and is-in. For instance in the statement, "Java is a Object-oriented language", the concepts 'Java' and 'object-oriented language' are captured using 'is-a' relationship. The is-in relationship is captured as a composition tree. For instance, "Object-oriented Languages" is contained within the concept of "Programming Languages".

Concept Extraction: It is the process of extracting the concepts in a given piece of text, by comparing the main terms with the concept tree.

3.2 Solution Approach

The basic premise of our approach is that a piece of information is going to be of use to a user, only if it captures some interest of that user. In domain-specific scenarios, like an academic environment, the interests of a user get directly mapped to their relevant subject areas. Based on this assumption, we methodize our approach of detecting spam by using the following principles.

- When a message is posted to a user, its relevant concepts are extracted and compared with that of the concepts extracted from the user's profile, interests and navigation history. If the semantics derived from both sides do not match, then that message is a candidate for Spam evaluation.
- When a message is posted in other resources, such as a common forum, the concepts involved in the message are compared against the subject domain and particularly with those concepts related with that resource and if they do not show any similarity then the message is a candidate for Spam evaluation.
- A statistical measure is also considered for detection of spam. If a message gets repeated across several resources within a short span of time, then that message is a candidate for Spam evaluation. Usually spammers would employ spam robots or scripts to do automatic postings.

3.3 Algorithm for Spam Detection

- **Input:** Message M, Concept Model of Resource R where message has been posted M_R.
- **Output:** Spam Hint Y/N.
- **Assumptions: a)** Presence of concept extraction algorithm.**b)** Presence of concept and containment trees.**c)** Resource R has been properly modeled.
- **Algorithm:**
  - Extract concepts from M using concept extraction algorithm and store in an array C_M
  - Extract concepts from M_R and store it an array C_R.
  - Find equivalent concepts corresponding to each concept in C_M and C_R from concept tree and add them to respective arrays.
  - Find the term frequency of occurrence of each concept in C_R and store it as weight-age against each concept.
  - Divide the concept array C_R in 3 parts with respect to the weight-age as highly probable C_{HP}, mid probable C_{MP} and low probable C_{LP} concepts.
  - Analyze M for relevant concepts by computing: $I_H = n(C_M \cap C_{HP})/ n(C_M)$, $I_M = (C_M \cap C_{MP})/ n(C_M)$, $I_L = (C_M \cap C_{LP})/ n(C_M)$
  - Define a Spam filter: = $(I_H < 0.5)$ AND $(I_M < 0.65)$ AND $(I_L < 0.8)$
- **Reason and Consideration:** We have categorized the matching concepts into 3 parts and fixed threshold for different groups (heuristics). They may vary depending on the subject domain, and the environment, with minor deviations.
- We have also defined a Spam filter that could be tweaked to suit different requirements, say a very strict filtering, or medium filtering.

4 Implementation and Results

A prototype of an academic networking website, with interactive features for its users was used for validating our approach. A messaging system and a discussion forum were built within the prototype system. Messaging was used by the participants for communicating among themselves, sharing information and resources. To test our approach, few spam messages were generated and posted to several participants, through scripts, and also to test further, certain spam messages were disguised as system notifications and posted.

Our proposed method was able to successfully detect and classify the spam posted to users in a large number of cases. Even in the case of spam disguised as system notification, the success ratio was fairly good, as the technique was concept based. Also upon activation of our method, logs were analyzed to find 'false' detection in participant messages, and system generated notifications, and though there were instances of false detection, they were very few and far between, and could be attributed to the lesser depth of concepts in concept tree.

In case of discussion forum, participants were split into groups, and each group had a lead, who kick-started the discussion with a lead question, and other participants in the group responded and took the discussions further. As the discussions progressed, 50 different spam messages were generated and put up for testing both manually and through scripts. Our proposed spam detection method was able to successfully discover and label most of the spam put up through scripts. In case of manual posting of spam, a mix of concepts related to the discussions were used, to fool the detection methodology. However even in such cases, the contextual spam detection was able to successfully detect 80% of them, and the technique of weighted concepts helped. The technique failed only, when the spam message was fairly large and contained many concepts relating to the conversation. The detection methodology was also evaluated for "false" detection and in this case of forums, it was considerably nil.

5 Conclusion

We described a method based on contextual strategies to detect spam in academic portals or sites. Application of this method detects the presence of spam more accurately, and effectively. We made the assumption that the resources in the system have been well defined and modeled. We also assumed the existence of ontology in terms of concept tree with simple relationships that defines the domain in which the academic portal operates.

This methodology could be easily extended to other domain-specific vertical portals, though the challenge would be in extending it to a generic and interactive information system.

References

1. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A Bayesian Approach to Filtering Junk E-Mail. In: Learning for Text Categorization: Papers from the 1998 Workshop (1998)
2. Cournane, A., Hunt, R.: An analysis of the tools used for the generation and prevention of spam. Computers & Security 23, 154–166 (2004)
3. Gyongyi, Z., Garcia-Molina, H.: Web spam taxonomy. In: Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web, Chiba, Japan (2005)
4. So Young, P., Jeong Tae, K., Shin Gak, K.: Analysis of applicability of traditional spam regulations to VoIP spam. In: The 8th International Conference on Advanced Communication Technology, ICACT 2006, pp. 3–1217 (2006)
5. Nagamalai, D., Dhinakaran, B.C., Lee, J.K.: An in-depth Analysis of Spam and Spammers. International Journal of Security and its Applications 2(2) (April 2008)
6. LinkedIn, `http://www.linkedin.com`
7. CiteULike, `http://www.citeulike.org`
8. ArnetMiner, `http://www.arnetminer.org`
9. Spam in Blogs, `http://en.wikipedia.org/wiki/Spam_in_blogs`
10. Hayati, P., Potdar, V., Talevski, A., Firoozeh, N., Sarenche, S., Yeganeh, E.A.: Definition of spam 2.0: New spamming boom. In: Digital Ecosystem and Technologies (DEST). IEEE Computer Society, Dubai (2010)

11. Hayati, P., Potdar, V.: Toward spam 2.0: An evaluation of web 2.0 anti-spam meth. In: 7th IEEE International Conference on Industrial Informatics, pp. 875–880. IEEE Computer Society, Cardi (2009)
12. Heymann, P., Koutrika, G., Garcia-Molina, H.: Fighting spam on social web sites: A survey of approaches and future challenges. IEEE Internet Computing 11, 36–45 (2007)
13. Benevenuto, F., Rodrigues, T., Almeida, V.A.F., Almeida, J.M., Goncalves, M.A., Ross, K.W.: Video pollution on the web. First Monday 4 (2010)
14. Bhattarai, A., Rus, V., Dasgupta, D.: Characterizing comment spam in the blogosphere through content analysis. In: IEEE Symposium on Computational Intelligence in Cyber Security (CICS), pp. 37–44. IEEE Computer Society Press, Nashville (2009)
15. Dawei, Y., Davison Brian, D., Zhenzhen, X., Liangjie, H., April, K., Lynne, E.: Detection of harassment on web 2.0. In: CAW2.0 Workshop at WWW 2009 (2009)
16. Dhinakaran, B.C., Nagamalai, D., Lee, J.-K.: Bayesian Approach Based Comment Spam Defending Tool. In: Park, J.H., Chen, H.-H., Atiquzzaman, M., Lee, C., Kim, T.-h., Yeo, S.-S. (eds.) ISA 2009. LNCS, vol. 5576, pp. 578–587. Springer, Heidelberg (2009)
17. Niu, Y., Wang, Y.-M., Chen, H., Ma, M., Hsu, F.: A quantitative study of forum spamming using context-based analysis. In: 14th Annual Network and Distributed System Security Symposium (NDSS), San Diego, CA, pp. 79–92 (2007)
18. Shin, Y., Gupta, M., Myers, S.: Prevalence and mitigation of forum spamming. In: IEEE INFOCOM. IEEE Computer Society (2011)
19. Social Networking Spam, http://en.wikipedia.org/wiki/Social_networking_spam
20. McFedries, P.: Technically Speaking: Slicing the Ham from the Spam. IEEE Spectrum 41, 72 (2004)
21. Han, S., Ahn, Y., Moon, S., Jeong, H.: Collaborative blog spam filtering using adaptive percolation search. In: WWW 2006 Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2006)
22. Narisawa, K., Yamada, Y., Ikeda, D., Takeda, M.: Detecting blog spam using the vocabulary size of all substrings in their copies. In: WWE 2006 3rd Annual Workshop on the Weblogging Ecosystem, Edinburgh, Scotland (2006)
23. Yu-Ru, L., Hari, S., Yun, C., Junichi, T., Belle, L.T.: Splog detection using self-similarity analysis on blog temporal dynamics. In: Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web. ACM, Banff (2007)
24. Zinman, A., Donath, J.: Is Britney Spears spam. In: Fourth Conference on Email and Anti-Spam Mountain View, California (2007)
25. Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., Zhang, C., Ross, K.: Identifying Video Spammers in Online Social Networks. In: AIRWeb 2008, Beijing, China (2008)
26. Georgia, K., Frans Adjie, E., Zolt, G.n, ngyi, Paul, H., Hector, G.-M.: Combating spam in tagging systems. In: Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web, ACM, Banff (2007)
27. Sureka, A.: Mining User Comment Activity for Detecting Forum Spammers in YouTube. In: The Proceedings of 20th WWW Conference, Hyderabad (2011)
28. Rajendran, B.: Socio-Contextual Filters for Discovering Similar Knowledge-Gathering Tasks in Generic Information Systems. In: Yang, C.C., Chen, H., Chau, M., Chang, K., Lang, S.-D., Chen, P.S., Hsieh, R., Zeng, D., Wang, F.-Y., Carley, K.M., Mao, W., Zhan, J. (eds.) ISI Workshops 2008. LNCS, vol. 5075, pp. 384–389. Springer, Heidelberg (2008)
29. Rajendran, B., Iyakutti, K.: Socio-contextual Network Mining for User Assistance in Web-based Knowledge Gathering Tasks. In: Memon, N., Alhajj, R. (eds.) From Sociology to Computing in Social Networks: Theory, Foundations and Applications. Lecture Notes in Social Networks, vol. 1, pp. 81–93. Springer, Wien (2010)