Journal of Computer Science 10 (10): 2135-2140, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.2135.2140 Published Online 10 (10) 2014 (http://www.thescipub.com/jcs.toc)

A STUDY OF SPAM DETECTION ALGORITHM ON SOCIAL MEDIA NETWORKS

[1]Saini Jacob Soman and [2]S. Murugappan

[1]Department of Computer Science and Engineering, Sathyabama University, India
[2]Department of Computer Science and Engineering, Annamalai University, India

Received 2013-11-21; Revised 2014-05-12; Accepted 2014-07-07

ABSTRACT

In today's world, the issue of identifying spammers has received increasing attention because of its practical relevance in the field of social network analysis. The growing popularity of social networking sites has made them prime targets for spammers. By allowing users to publicize and share their independently generated content, online social networks become susceptible to different types of malicious and opportunistic user actions. Social network community users are fed with irrelevant information while surfing, due to spammer's activity. Spam pervades any information system such as e-mail or web, social, blog or reviews platform. Therefore, this study attempts to review various spam detection frameworks which deals about the detection and elimination of spams in various sources.

Keywords: Spam Detection, Spam Analysis, Feature Extraction

1. INTRODUCTION

Social networks such as Face book, MySpace, LinkedIn, Friendster (Danah, 2004) and Tickle have millions of members who use them for both social and business networking. Due to the astonishing amount of information on web, users follow the way of searching useful web pages by querying search engines. Given a query, a search engine identifies the relevant pages on the web and presents the users with the links to such pages. Spammers try to increase the page rank of the target web page in search results by Search Engine Optimization (SEO), the injection of artificially created pages into the web in order to influence the results from search engines to drive traffic to certain pages for fun or profit. Initially, spams are introduced in mails. Later, this has been extended to Social networks. On the other hand, in E-mail system spammer sends unsolicited bulk email to users by redirecting them to irrelevant websites. The success of delivered attacks is dependent almost entirely upon the click-through rate of the email. If the target does not click on the malicious link presented in the email, then the attack usually fails. To improve click-through rates, many techniques exist such as: Hiding the destination of hyperlinks, falsifying header information and creative use of images (Markus and Ratkiewicz, 2006; Alex and Jakobsson, 2007).

Email messages also takes advantage of some shared context among friends on a social network such as celebrations of birthday functions, residing in the same home town, or common events participation. This shared context dramatically increase email authenticity, filters and increasing the click-through rate for spam that contains advertisements, installs malicious software, or solicits sensitive personal information (Takeda and Takasu, 2008; Kamaliha *et al*., 2008). But in the content of blog platforms, spammer post irrelevant comments for an already existing post. They focus on several kinds of spam promotion such as, splog (the whole blog is used to promote a product or service), comment spam (comments promoting services with no relation to the blog topic) and trace back spam (spam that takes advantage of the trace back ping feature of popular blogs to get links from them). This study primarily focusses on the survey of the literature which deals with the comment spams in blog. Since comments are typically

Corresponding Author: Saini Jacob Soman, Department of Computer Science and Engineering, Sathyabama University, India

short by nature. Comment spam is essentially link spam originating from comments and responses added to web pages which support dynamic user editing. As a result of the presence of spammers in a network, there is a decrease in both the quality of the service and the value of the data set as a representation of a real social phenomenon. With the help of the extracted features it is possible for identifying spammers from the legitimate one. Various machine learning, supervised and unsupervised methods have been used in the literature for classification of these spams. A study of various spam detection algorithms has been dealt thoroughly in this study.

2. RELATED WORK

Initially, certain researchers concentrated on the development of Honey pots to detect spams. To detect spams, (Webb *et al*., 2008) dealt with automatic collection of deceptive spam profiles in social network communities based on anonymous behavior of user by using social honey pots. This created unique user profiles with personal information like age, gender, date of birth and geographic features like locality and deployed in MySpace community. Spammer follows one of the strategy such as being active on web for longer time period and sending friend request. The honey profile monitors spammers behavior by assigning bots. Once the spammers sends friend request the bots stores spammer profile and crawls through the web pages to identify the target page where advertisements originated.

The spammer places woman's image with a link in the "About Me" section in its profile and the honey profile bots crawls through the link, parses the profile, extracts its URL and stores the spammers profile in the spam list. URL does not redirect at times during crawling process and "Redirection Detection Algorithm" is executed to parse the web page and extract redirection URL to access it with the motive of finding the source account. Also, he proposed a "Shingling Algorithm" which verifies the collected spam profile for content duplication like URL, image, comments and to accurately cluster spam and non-spam profile based on the features. In this way, he eliminated Spams.

Another researcher named (Gianluca *et al*., 2010) used social honey pots to construct honey profiles manually with features like age, gender, dob, name, surname etc. Here, the honey profiles have been assigned to three different social network communities (Myspace, Facebook, Twitter). It considered friend request as well as the message (wall post, status updates) received from spammers and validate with honey pots. The identified spam accounts with the help of spam bots share common traits which has been formalized as features in their honey pots (first feature, URL ratio, message similarity, friend choice, message sent, friend number etc). The classifier namely Weka framework with a random forest algorithm has been used to classify spammers for best accuracy. Similarly during spam campaign the spam bots were clustered based on spam profiles using naïve Bayesian classifier to advertise same page during message content observation.

Similarly, (Lee *et al*., 2010) dealt with the social spam detection which has become tedious in social media nowadays. Here Social honey pots are deployed after its construction based on the features such as number of friends, text on profile, age, etc. Here, both legitimate and spam profiles have been used as initial training set and Support Vector Machine has been used for classification. An inspector has been assigned to validate the quality of extracted spam candidates using "Learned classifier" and provide feedback to spam classifier for correct prediction in future. In this study three research challenges have been addressed. Initially, it validates whether the honey pot approach is capable of collecting profiles with low false positives, next to that it addresses whether the users are correctly predicted and finally it evaluates the effectiveness of fighting against new and emerging spam attacks.

The first challenge is proved using automatic classifier which groups the spammers accurately. The second one considers demographic features for training the classifier using 10-fold cross validation. It has been tested in MySpace using Meta classifier. In twitter it used Bigram model for classification along with the preprocessing steps. Finally, post filters has been used to check the links and remove the spam label by applying "Support Vector Machine" for correct prediction. They also proposed that in future, Clique based social honey pots can be applied with many honey profiles against many social network communities.

Next to honey pots, Spammers have been identified in the literature by analyzing content and link based features of web pages. In this context, (Sreenivasan and Lakshmipathi, 2013) has performed spam detection in social media by considering content and link based features.

Web spam misleads search engines to provide high page rank to pages of no importance by manipulating link and contents of the page. Here, to identify web spam, Kullback-leiblerence techniques are used to find

the difference between source page features (anchor text, page title, meta tags) and target page features (recovery degree, incoming links, broken links etc). Therefore, three unsupervised models were considered and compared for web spam detection. As a first one, Hidden Markov model has been used which captures different browsing patterns, jump between pages by typing URL's or by opening multiple windows. The features mentioned above were given as input to HMM and it is not visible to the user. As a result, a link is categorized as spam or non spam based on how frequently a browser moves from one page to another.

Second method uses "Self Organizing maps" a neural model to classify training data (Web links) without human intervention. It classifies each web link as either spam or non spam link. One more method called Adaptive Resonance Theory has also been used to clarify a link as either spam or not.

Another work in the literature (Karthick *et al*., 2011) has dealt with the detection of link spam through pages linked by hyperlink that are semantically related. Here, Qualified Link Analysis (QLA) has been performed. The relation existing between the source page and target page is calculated by extracting features of those two pages from web link and compared with the contents extracted from these pages. In QLA, the nepotistic links are identified by extracting URL, anchor text and cached page of the analyzed link stored in the search engines. During query generation, once the page is available with search engines, this result has been compared together with the page features for easy prediction of spam and non spam links.

In this study, QLA has been combined with language model detection for better prediction of spams. In Language model detection, the KL divergence technique has been used to calculate the difference between the information of the source pages with the content extracted from the link. Once matched, it is clustered as non spam and vice versa. Here, the result of LM detection, QLA along with pre trained link and content features lead to accurate classification and detection.

Qureshi *et al*. (2011) handled the problem of eliminating the existence of irrelevant blogs while searching for a general query in web. The objective is to promote relevancy in ordering of blogs and to remove irrelevant blogs from top search results. The presence of irrelevancy is not because of spam, but is due to inappropriate classification for a topic against a query. This approach uses both content and link structure features for detection. The content features calculate the cosine similarity between blog text and blog post title while searching for a particular blog. It has been proved that a co-relation exist between the above two features with which the spammer activity is detected based on the degree of similarity. This detection achieved a precision of 1.0 and recall of 0.87.

The blog link structure feature finds spammer activity by decoupling between two classes (duplicate and unique links) up to three hop counts. Spammers always move within closed group rather than with other blogosphere. The duplicate links are identified and removed.

Wang and Lin (2011) focused on comment spams with hyperlinks. The similarity between the content of page for a post to the link it points to has been compared to identify spam. Here, the collected blogs are preprocessed which finds the stop word ratio that is found to be less in spammers post. The contents are extracted from the post and are sorted where "Jaccard and Dice's " co-efficient is calculated which provides the degree of overlapping between words. The degree of overlapping is used for calculating inter comment similarity for a comment with respect to a post. Analysis of content features like inter comment similarity and post comment similarity along with the non-content features like link number, comment length, stop words showed better results in identifying spam links.

Next to this, comment based spams have also been discussed here. Archana *et al*. (2009) has dealt with the spam that gets penetrated in the form of comments in Blog. A blog is a type of web content which contains a sequence of periodic user comments and opinions for a particular topic. Here, spam comment is an irrelevant response received for a blog post in the form of a comment. This comments are analyzed using supervised and semi supervised methods. Analysis considers various features to identify spams. They are listed below: The post similarity feature has been used to find the relevancy between the post and the comment. Word Net tool has been used to spot out the word duplication features. Word duplication feature identifies the redundant words in comments and it is found to be higher for spam comments and low for genuine comments. Anchor text feature counts the number of links exists for a comment and predicts that the spammers are the one having higher count. Noun concentration feature has been used to extract comments and part of speech tags from the sentences. In that, the legitimate users have low noun concentration.

Stop word ratio feature consider sentences with a finishing point where spammers have less stop word

ratio. Number of sentence feature counts the number of sentences exists in a comment and is found to be higher for spammers. Spam similarity feature checks for the presence of spam words listed and categorize it. The words identified as spam after preprocessing were assigned a weightage and the contents which falls above the threshold are detected as spam comments. Here, a supervised learning method (Naïve Baye's classifier) has been used along with pre classified training data for labeling a comment as spam and non spam. One more unsupervised method directly classifies the comments based on the expert specified threshold.

Interestingly in literature, works have been carried out for book spammers also. Sakakura *et al*. (2012) deals with bookmark spammers who create bookmark entries for the target web resource which contains advertisements or inappropriate contents thereby creating hyperlinks to increase search result ranking in a search engine system. Spammer may also create many social bookmark accounts to increase the ranking for that web resource. Therefore, user accounts must be clustered based on the similarity between set of bookmarks to a particular website or web resource and not based on the contents. Here in this study, data preprocessing is done by clustering bookmarks by extracting web site URL from the raw URL since spammer may create different bookmark entry for same URL.

Here, the similarity based on raw URL (which is the ratio of number of common URL'S to total number of URL'S contained in the bookmarks of two accounts) has been considered and the similarity based on site URL without duplicates (which is the ratio of number of common site URL'S to the total number of all URL's in both the accounts) and the similarity based on site URL with duplicates (Weight of the sites based on the number of bookmarks common to the user accounts) has been calculated. The agglomerative hierarichal clustering of accounts has been made based on one of the above mentioned similarities. The cluster which is large and having higher cohesion is categorized as an intensive bookmark account spammer. This study achieves a precision of 100%.

Yang and Chen (2012) this study deals with online detection of SMS spam's using Naïve Bayesian classifier, which considers both content and social network SMS features. The SMS social network is constructed from the historical data collected over a period with the help of telecom operator. The content features are extracted that are presented in vector space and the weights are assigned to the vector obtained using term frequency function. The feature selection methodologies like information gain and odd ratio has been used for selecting words from SMS with which class dependency and class particularity are found for clustering "content based features". Features on social network tries to extract both the sending behavior of mobile users and closeness for categorizing spammer and legitimate user. Bloom filter is used to test the membership between sender and receiver for removing spammer's relationship. Naïve Bayesian classifier has been used for classifying users as legitimate or spam using the above features.

Ravindran *et al*. (2010) deals with the problem of tag recommendation face which contains popular tags for particular bookmarks based on user feedback and to filter spam posts. In this problem, Spammer may increase the frequency of a particular tag and the system may suggest those tags which have higher frequency to the user. To eliminate this problem, this study uses "frequency move to set "model to choose a set of tags suggested by user for a bookmark. To find whether a tag is popular or not for placing it in the suggestion set, the tag feature like simple vocabulary similarity has been considered. The suggestion set which is kept updated is measured using the stagnation rate and unpopular tags are removed randomly from the set. The decision tree classifier has been used here to classify tags as spam and non-spam. The accuracy obtained in this approach is about 93.57%.

Ariaeinejad and Sadeghian (2011) deals with detecting email spam in an email system by considering plain text alone for categorizing a mail as spam or ham. The common words in spam and ham emails are eliminated and stored in white list. The collected words are parsed by removing unwanted spaces and other signs among the words. The parsed words are compared with white list and common words are eliminated. The cleaned words are checked for making decision using "Jaro-Wrinkler" technique. Here, a fuzzy map is constructed as a two dimension using an interval type and 2 fuzzy methods have been used which represents distance of each word in email with closed similarity in dictionaries as a horizontal vector and represents weight of the words in dictionary as an vertical vector. Third dimension considers importance of a word in an email and its frequency which is identified using term frequency inverse document frequency technique. Here, Email has been categorized into spam, ham and uncertain

zone using fuzzy-C means clustering. Later, the words are updated consistently for correct prediction.

Another work reported by (Ishida, 2009) deals with detection of spam blogs and keywords mutually by its co-occurrence in the cluster. He employed shared interest algorithm for the blogs collected. This algorithm constructs a bi-partite graph between blogs and low frequency keywords from which clusters of varying size has been formed. The spam score for each cluster is calculated and are ranked by multiplying number of blogs and keywords in the cluster. The spam blogs with highest score is considered as spam seed and is stored in a list. A threshold is set manually for the ranked spam blogs and keywords. Those which exceeds the threshold has been detected as spam blog, spam keywords are removed from the list. This approach provides mutual detection and thereby reducing the filtering cost and the words are kept updated.

3. CONCLUSION

This survey has presented various approaches which could identify or detect spams in the social network by extracting necessary information from web pages. Many researchers worked on Honey pot profiles, whereas a few people worked on identifying spam links. Even works have been carried out on email and SMS spams. But still this area remains in its infant stage and more number of spam detection algorithms need to be devised for social media networks.

4. REFERENCES

Alex, T. and M. Jakobsson, 2007. Deceit and deception: A large user study of phishing. Pennsylvania State University.

Archana, B., V. Rus and D. Dasgupta, 2009. Characterizing comment spam in the blogosphere through content analysis. Proceedings of the IEEE Symposium on Computational Intelligence in Cyber Security, Mar. 3-3, IEEE Xplore Press, Nashville, TN, pp: 37-44. DOI: 10.1109/CICYBS.2009.4925088

Ariaeinejad, R. and A. Sadeghian, 2011. Spam detection system: A new approach based on interval type-2 fuzzy sets. Proceedings of the 24th Canadian Conference on Electrical and Computer Engineering, May 8-11, Niagara Falls, pp: 379-384. DOI: 10.1109/CCECE.2011.6030477

Danah, M.B., 2004, Friendster and publicly articulated social networking. Proceedings of Extended Abstracts on Human Factors in Computing Systems, (CHI '04), Vienna, Austria, pp: 1279-1282. DOI: 10.1145/985921.986043

Gianluca, S., C. Kruegel and G. Vigna, 2010. Detecting spammers on social networks. Proceedings of the Annual Computer Security Applications Conference, Dec. 6-10, New York, pp: 1-9. DOI: 10.1145/1920261.1920263

Ishida, K., 2009. Mutual detection between spam blogs and keywords based on cooccurrence cluster seed. Proceedings of the 1st International Conference on Networked Digital Technologies, Jul. 28-31, IEEE Xplore Press, Ostrava, pp: 8-13. DOI: 10.1109/NDT.2009.5272171

Wang, J.H. and M.S. Lin, 2011. Using Inter-comment similarity for comment spam detection in chinese blogs. Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, Jul. 25-27, IEEE Xplore Press, Kaohsiung, pp:189-194. DOI: 10.1109/ASONAM.2011.49

Kamaliha, E., F. Riahi, V. Qazvinian and J. Adibi, 2008. Characterizing network motifs to identify spam comments. Proceedings of the IEEE International Conference on Data Mining Workshops, Dec. 15-19, IEEE Xplore Press, Pisa, pp: 919-928. DOI: 10.1109/ICDMW.2008.72

Karthick, K., V. Sathiya and J. Pugalendiran, 2011. Detecting nepotistic links based on qualified link analysis and language models. Int. J. Comput. Trends Tech.

Lee, K., J. Caverlee and S. Webb, 2010. Uncovering social spammers: Social honeypots + machine learning. Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 19-23, New York, pp: 435-442. DOI: 10.1145/1835449.1835522

Markus, J. and J. Ratkiewicz, 2006. Designing ethical phishing experiments: A study of (ROT13) rOnl query features, Proceedings of the 15th International Conference on World Wide Web, May 22-26, IEEE Xplore Press, New York, pp: 513-522. DOI: 10.1145/1135777.1135853

Qureshi, M.A., A. Younus, N. Touheed, M.S. Qureshi and M. Saeed, 2011. Discovering irrelevance in the blogosphere through blog search. Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, Jul. 25-27, IEEE Xplore Press, Kaohsiung, pp: 457-460. DOI: 10.1109/ASONAM.2011.84

Ravindran, P.P., A. Mishra, P. Kesavan and S. Mohanavalli, 2010. Randomized tag recommendation in social networks and classification of spam posts. Proceedings of the IEEE International Workshop on Business Applications of Social Network Analysis, Dec. 15-15, IEEE Xplore Press, Bangalore, pp: 1-6. DOI: 10.1109/BASNA.2010.5730294

Sakakura, Y., T. Amagasa and H. Kitagawa, 2012. Detecting social bookmark spams using multiple user accounts. Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, (NAM' 12), Washington, pp: 1153-1158. DOI: 10.1109/ASONAM.2012.199

Sreenivasan, S. and B. Lakshmipathi, 2013. An unsupervised model to detect web spam based on qualified link analysis and language models. Int. J. Comput. Applic., 63: 33-37. DOI: 10.5120/10455-5163

Takeda, T. and A. Takasu, 2008. A splog filtering method based on string copy detection. Proceedings of the 1st International Conference on Applications of Digital Information and Web Technologies, Aug. 4-6, IEEE Xplore Press, Ostrava, pp: 543-548. DOI: 10.1109/ICADIWT.2008.4664407

Yang, Y. and Y. Chen, 2012. A novel content based and social network aided online spam short message filter. Proceedings of 10th World Congress on Intelligent Control and Automation, Jul. 6-8, IEEE Xplore Press, Beijing, pp: 444-449. DOI: 10.1109/WCICA.2012.6357916

Webb, S., J. Caverlee and C. Pu, 2008. Social honeypots: Making friends with a spammer near you. Paper Presented Meeting CEAS.