Toward free spam social networks: detecting and tracking spammers in Twitter

Mahdi Washha, Florence Sèdes

Toward free spam social networks: detecting and tracking spammers in Twitter

Mahdi Washha , Florence Sèdes

Institut de Recherche en Informatique de Toulouse
Avenue de l'étudiant
31400 Toulouse

Mahdi.Washha@irit.fr

MOTS-CLÉS : *Réseaux Sociaux, Spam, Utilisateurs Légitimes, Machine Learning, Temps.*

KEYWORDS: *Social Networks, Spam, Legitimate Users, Machine Learning, Time.*

ENCADREMENT. *Florence Sedes (PR)*

1. Context

Social networks are now so well established with playing an important role at communication level between people. However, the propagation of noisy information, so-called "spam", in social networks is growing up daily at unusual rates where unethical goals stand behind publishing spam information. The spreading of spam affects negatively in different aspects, summarized in : (i) polluting real-time search ; (ii) interfering on statistics computed by mining tools ; (iii) consuming significant resources from both humans and systems ; (iv) decreasing the performance of search engines that use explicitly social signals ; and (v) violating the privacy of user's which occurs because of viruses and phishing methods. Thus, this research aims at finding out solutions suitable for the applications that take social networks as a source of information. Also, we target to overcome the existing limitations in the current state of art solutions.

2. State of art

2.1. *Spam definition and spammers' goals*

The first appearance of spam problem has been in electronic messaging systems. Researchers have defined the spam as unsolicited and undesired messages sent to the systems users by unethical individuals, so called "spammers" (Qaroush *et al.*, 2012). However, spammers have extended their targets to include social networks area because of the huge volume of users, estimated at hundreds of millions.

Spammers have a set of goals, e.g. spreading advertisements to generate sales, disseminating pornography, viruses and phishing. They achieve their spamming behavior

through exploiting the services that social networks provide such as hashtags and shorten URLs. Also, spammers leverage the availability of free APIs provided by social networks to automate the publishing of spam.

2.2. *Implicit solutions provided by social networks*

The social networks administrators have attempted to break up the spam phenomenon. Their attempts are centered around using users' reports, and defining general rules. Exploiting the users' reports can be viewed as a kind of manual collaboration. However, such an attempt needs a manual effort from both administrators (review sent reports) and users (send report about accounts). The rules concept can contribute in detecting spammers by suspending permanently upon violating the rules defined. However, the current social networks like Twitter use general rules such that spammers can easily bypass them.

2.3. *Automated spam detection solutions*

The limitations in the solutions of social networks have motivated researchers to propose more powerful methods. We classify their methods into two approaches based on the degree of automation in detecting spam content or spammers : (i) machine learning approach ; (ii) social honeypot approach. In this paper, we just mention the related work of machine learning approach as a fully automated one, because of the inefficiency of social honeypot based solutions.

The existing solutions use the concept of features combined with learning algorithms to automate the detection process. However, as main differences between the solutions, machine learning techniques have been employed at three levels of detection : (i) post-level ; (ii) user-level ; and (iii) campaign-level.

Post-level. At this level, Martinez-Romo and Araujo (Martinez-Romo *et al.*, 2013) applied probabilistic language models to determine the topic of the considered post. Then, they label the post as spam when it is too diverging from the potential topic. Benevenuto (Benevenuto *et al.*, 2010) identified spam tweet through extracting a set of features like number of words from each tweet individually. They applied then the Support Vector Machine (SVM) learning algorithm on manually annotated data-set to have a binary classifier.

User-level. The work investigated in (Benevenuto *et al.*, 2010) proposed features associated to the account, including number of followers, number of friends, similarity between tweets posted or re-tweeted by the user, ratio of URLs in tweets. Much more complex features related to the graph theory have been extracted at the user level as well. For example, (Yang *et al.*, 2011) examined the relation between users by using graph metrics as features, including local clustering, node betweenness, and bi-directional relation ratio. Song, Lee, and Kim (Song *et al.*, 2011) studied some relational features such as the distance and the connectivity between sender and receiver(s) of a given message. In the same level, user profiling methods and community clustering algorithms (Mezghani *et al.*, 2015) can contribute better than using abstract graph metrics. However, the time complexity issue in such methods prevents to exploit them in fighting spammers.

Campaign-level. Chu et al. (Chu *et al.*, 2012) clustered users' accounts based on the URLs retrieved from their posted tweets. Then, a set of features from the clustered accounts were extracted to be incorporated in identifying spam campaign. Indeed, this level is not applicable to detect individual spammers as well as spammers can lunch uncorrelated small campaigns to avoid detection.

3. Problem statement

State of art solutions are still not robust enough to adapt the changes of spammers' tactics in propagating spam information. More precisely, post-level methods are an inefficient solution to identify spammers because one post does not provide "informative" information to distinguish the spammers' behavior from the legitimate users' behavior. Therefore, relying on such way in detecting spam content is not a practical solution.

The user-level methods have critical limitations and major drawbacks derived from using features that are easy to manipulate. Thus, spammers can avoid the detection when using such features. As an example, the number of followers (i.e. the accounts that follow a user) is mainly used in detecting spammers as a feature. Based on the state of art conclusions, the small number of followers has high probability for being spammer. However, such a number can be easily increased by creating a huge number of recent accounts with letting each account to follow each other, and thus keep away from being classified as spammer. Also, although of graph features give high accuracy, but they remain incompatible with real time filtering. This incompatibility is because of the time complexity issue of graph metrics and the necessity to fetch all users of social network to get the metrics values.

At the campaign level, working at this level is effective to detect big campaigns only, not for individual spammers. Furthermore, spammers are intelligent enough to design spam campaigns such that the correlations between accounts are not detectable in easy ways.

With the negative impacts of spam on social networks and the limitations in the current state of art methods, our research problem is summarized in answering the following research questions :

– To what extent can the creation date of the account and the posting date of the post contribute in detecting spammers, as unmodifiable attributes by social networks users ?

– Is it possible to have few and robust features suitable for real-time spammers detection ?

– Instead of searching in the all users of social networks, is it possible to define a heuristic function such that it can predict the names of spam accounts, as a way to locate and track spammers ?

4. Work in progress

In addressing the problem of identifying spammers, we propose a design of new generic features suitable for real-time filtering. Our features are distributed between statistical features, behavioral features. The statistical ones incorporate explicitly both the time of posting tweet and the creation date of user's account. We choose the time and date attributes because they are unmodifiable attributes overtime. Indeed, this explains our motivation in leveraging them. The behavioral features can catch any potential posting behavior

similarity over time between different instances (e.g. hashtags) available in the user's posts (e.g. correlation between different hashtags (#h1 and #h2) posting time). Our features are inspired by the following hypotheses : (i) spammers tend to create recent accounts ; (ii) spam accounts follow each other to boost up some critical attributes such as number of followers ; (iii) spammers have systematic and defined posting pattern.

We validate the generic features proposed using Twitter social network as an example. In doing so, we crawled a data-set consisting of 7,100 users from Twitter social network. Then, we annotated manually the crawled users by assigning a label (spammer or non-spammer) for each user. This forms a data set available as ground truth for other assessments. Using the annotated data-set, we employed machine learning algorithms to build a binary classifier using the crawled data-set and features designed. According to the experiments, our new features are able to classify correctly the majority of spammers with a detection rate higher than 93% when applying Random Forest as a classification algorithm. The results obtained outperform by 6% the detection rate of 70 state of art features proposed in (Benevenuto *et al.*, 2010). This work is going to be published as one contribution in spam detection area.

5. Future works and perspectives

The next steps will be dedicated to finding a simple and fast method to locate and track spammers in a social network. In doing so, we plan to study the correlation between spammers' names and the current occurring events in social networks. This may help in predicting the name of spam accounts, which forms an entry point to locate spammers. We motivate this correlation study based on a true observation found in our crawled data-set. The observation found states that spammers leverage often trending events to publish spam information using accounts having names inspired by the content and description of events.

6. References

Benevenuto F., Magno G., Rodrigues T., Almeida V., « Detecting spammers on twitter », *In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS*, 2010.

Chu Z., Widjaja I., Wang H., « Detecting social spam campaigns on twitter », *Applied Cryptography and Network Security*, Springer, p. 455–472, 2012.

Martinez-Romo J., Araujo L., « Detecting malicious tweets in trending topics using a statistical analysis of language », *Expert Systems with Applications*, vol. 40, n° 8, p. 2992–3000, 2013.

Mezghani M., On-at S., Peninou A., Canut M.-F., Zayani C. A., Amous I., Sedes F., *New Trends in Databases and Information Systems : ADBIS 2015, France, September 8-11, 2015. Proceedings*, Springer International Publishing, chapitre A Case Study on the Influence of the User Profile Enrichment on Buzz Propagation in Social Media : Experiments on Delicious, p. 567–577, 2015.

Qaroush A., Khater I. M., Washaha M., « Identifying spam e-mail based-on statistical header features and sender behavior », *Proceedings of the CUBE International Information Technology Conference*, ACM, p. 771–778, 2012.

Song J., Lee S., Kim J., « Spam filtering in twitter using sender-receiver relationship », *Recent Advances in Intrusion Detection*, Springer, p. 301–317, 2011.

Yang C., Harkreader R. C., Gu G., « Die Free or Live Hard ? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers », *Proceedings of the 14th International Conference on Recent Advances in Intrusion Detection*, RAID'11, Springer-Verlag, Berlin, Heidelberg, p. 318–337, 2011.