Detecting Spam from All Corners of the Web

Jonathan Cornwell
Florida State University, Florida
Jmc10@my.fsu.edu

Abstract

Spam is the use of electronic messaging systems to send unsolicited bulk messages, especially advertising, indiscriminately. Origionally this was mostly only a problem through e-mail applications, and much research has been done to detect and filter out these messages. But as the web continues to grow, spammers are finding new and more resourceful ways to spam throughout the web. Spammer's resources now range from e-mail and social networks to critiques and reviews. This paper reviews some different ways created or theorized to detect spammers from all corners of the web and introduces a way to combine these methods to form a new method which could potentially apply percantage based values to each of these methods an create an extremely accurate, easilty adaptable method of spam detection that would could potentially increase social media spam detection accuracy.

1 Introduction

With the development of the web ever increasing, people are more likely to express view and opinions all over the web with even more increasing ways of discussion, contact, and messaging. This papers specifically covers 4 areas: social networks, web pages, review sites, and deobfuscation. In each of these areas spammers have found ways to obscure and misuse intended purposes to promote their own agenda, resulting in a loss of a websites integrity, taking advantage of users, and all-around hate for spam in general. This paper proposes a hypothesis of combining multiple methods of spam detection for different areas across the web, applying percentage based values to each individual spam detection method, for a final combined value that should could be used to better detect social media spam, considering that social media spam uses all forms of spam at the same time.

1.1 Identifying Spammers in Social Networks

Researchers at the Hong Kong University of Science and Technology in Hong Kong have challenged themselves to discover new ways of finding spammers in social media[Zhu et al., 2012]. In social media sites, spammers create fake accounts and hijack regular users for personal gains, while e-mail spam can only be received through messages. These two types of spam differ greatly, with social media spammers being extreemely hard to detect because they have the ability to act like normal users. Social media sites depend on other users to report spammers, which detect spammers too late; they have already had their effect. A proposed "Supervised Matrix Factorization Method with Social Regularization" method has been developed for spammer detection in social networks that exploits both social activities as well as users' social relations in an innovative and highly scalable manner.

One of the problems with detecting social spam is the fact that possible social activites have more variaties than before and create a much larger feature space. Another is the addition of user privacy, where users just plain hate having anything scan the content of their social media, even anti-spammers. Finally, the behaviors of social media spam change rapidly, a system capable of capturing spam one week may not have this ability the next week. Spammers are creating new, more devious ways to create new, more evasive accounts, which is making social spam detection a great challenge for researchers.

The proposed technique involved creates a user set of spammers and non spammers combined with their social activities and relations. This set has a truth value for each user stating weather or not the user is a spammer or not. A matrix is then built for the users social activities. With two kinds of activities introduced – activities performed on his/her own social page, and interactive activities with other users, with the 2nd kind being majority. A user relation matrix is then introduced to state weather or not two users are friends or not. Using all this information a binary classifier is created, which predicts wheather a unlabeled user is a spammer or not. The most important thing to note from all this data is that a spammer's actions differs from his friends, no matter how smart the spammer is.

1.2 Identifying Web Page Spammers

Web page spam is an obscure form of spam that involves an entire web page being the spam itself. A user is often lead to these sites through varius methods involving re-direction, misleading links, and annoying pop-ups. Researchers at Dalian Univ. of Technology in Dalian, China have created a method of detecting this spam using a Trust-Distrust Rank (TDR) algorithm which propigates trust/distrust ranks from a set of seed (good/bad) pages[Zhang et al., 2011]. This has been used before, but very little work has been done to use this method successfully. This method specifically uses the fact that every web page has both a trustworthy side and an untrustworthy side, and two scores are assigned to each web page: T-Rank, which scores the trustworthiness of a page, and D-Rank, which scores how untrustworthy a page is. The TDR algorithm overcomes some of the disadvantages of existing trust/distrust propagation algorithms and makes full use of both good and bad seeds.

A page pointed to by multiple reputable pages is automatically consitered reputable, and a page pointing to many spam pages is automatically considered a spam page itself. Using this information, a TDR algorithm can act on an intial set of seed pages with preset values, which is a subset of a much larger set of pages. The resulting score values based of of each pages final T-Rank and D-Rank will tell if a page is spam or not.

1.3 Identifying Review Spam

Reseachers at Tsinghua University in Beijing, China are working on new ways, using machine learning methods, to detect a form of spam that has attacked without rebuttal, review spam[Li et al., 2011]. Review spam is a lage problem on product review sites where people can write faked reviews to promote their products or defame competitors' products. Little has gone into this research area before, with previous work only focusing on small heuristic rules, such as helpfulness voting, or rating deviation, limiting this tasks performance. The addition of machine learning methods to identify review spam has achieved significant improvements in comparison to the heuristic baselines. These researches propose a two-view semi-supervised method, call co-training, to exploit large amounts of unlabeled data based off of observations about review spammers' consistencies, especially the fact that review spammers review's are almost always spam.

1.4 Identifying Spam Deobfuscation

This subject area can be very misleading. Spam deobfuscation is a process to detect obfuscated words appearing in spam text, and to convert them back to the origional words for correct recognition[Lee et al., 2007]. Obfuscation is a form of spam mostly used in, but not limited to, e-mail spam. It's not an average type of e-mail spam though. Most e-mail spam is properly filtered using many forms of spam filtering technologies including content-based, and rule-based approaches. Spam deobfuscation was developed after these methods as a way to circumvent detection. By obfuscating words (such as turning the word Viagra into V.a | @ g r a), spam detectors can't detect this as spam and will make it to the inbox of unsuspecting users. Spam obfuscation can be categorized into five different ways: replacement, insertion, deletion, segmentation and mixed form. Deobfuscation has the ability to translate obfuscated words to their origional meaning, and apply normal e-mail spam filtering to them for correct results.

2 Combining Methods

The worst thing about social media spam is the fact that all of these forms of spam are used at the same time. We can use this to our advantage! As stated before, a social media spammer has two forms of spam, something that can be posted on their own page, and interactive activities with other users. Both of these types of activities are almost always web spam (links to other websites, or even redirection pages that force you to make your own link to the said website), or review spam (trying to get you to buy a product). And both of these types of spam are combined with obfuscation as a way to get around current web spam detection! Note: This is not inteded to detect the spam created by the actual website itself as a form of payed advertising.

The spam that we're trying to detect here is specifically spam spread through links. These links are combined with comments before the link – a combination of review and web page spam. Using the set from part 1.1 – you can use the algorithms discussed in parts 1.2 and 1.3 after using the method discussed in part 1.4 to deobfuscate any obfuscated words. This would result in two separate forms of spam detection to combine with the already proposed algorithm for social media spam detection. Applying percentage based values to these, with the origional form of spam detection having the highest base value, comes the proposed combined spam detection method.

References

[Zhu et al., 2012] Yin Zhu, Xiao Wang, Erheng Zhong, Nanthan N. Liu, He Li, and Qiang Yang. *Discovering Spammers in Social Networks.* Hong Kong University of Science and Technology, Hong Kong, 2012, www.aai.org/ocs/index.php/AAAI/AAAI12/paper/view/5073/5135.

[Zhang et al., 2011] Xianchao Zhang, You Wang, Nan Mou, and Wenxin Liang. *Propagating Both Trust and Distrust with Target Differentaition for Combating Web Spam*, School of Software Dalian Univ. of Technology, www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3600/4066.

[Li et al., 2011] Fangtao Li, Minlie Huang, and Yi Yang, Xiaoyan Zhu. *Learning to Identify Review Spam*,

Tsinghua University, Beijing, China. www.ijcai.org/papers11/Papers/IJCAI11-414.pdf.

[Lee et al., 2007] Seunghak Lee, Iryoung Jeong, and Seungjin Choi. *Dynamically Weighted Hidden Markov Model for Spam Deobfuscation*. POSTECH, Korea. www.ijcai.org/papers07/Papers/IJCAI07-406.pdf.