A Self-Supervised Approach to Comment Spam Detection Based on Content Analysis

A. Bhattarai, University of Memphis, USA
D. Dasgupta, University of Memphis, USA

ABSTRACT

This paper studies the problems and threats posed by a type of spam in the blogosphere, called blog comment spam. It explores the challenges introduced by comment spam, generalizing the analysis substantially to any other short text type spam. The authors analyze different high-level features of spam and legitimate comments based on the content of blog postings. The authors use these features to cluster data separately for each feature using K-Means clustering algorithm. The authors also use self-supervised learning, which could classify spam and legitimate comments automatically. Compared with existing solutions, this approach demonstrates more flexibility and adaptability to the environment, as it requires minimal human intervention. The preliminary evaluation of the proposed spam detection system shows promising results.

Keywords: Self-Supervised Learning, Spam Feature Extraction, Spam Filtering, Text Analysis, Text Characteristics

1 INTRODUCTION

Spam has evolved and expanded into different forms recently. Its initial root is the email domain. Blogs are becoming easy and popular targets for spammers due to the candidness with which people write their views and opinions. According to comScore (2009), the conversational media, which includes social networks and blogs, is the second highest growing category after games attracting around 75% web users with a growth rate of 7.1%. While this property has allowed it to increase its quantity tremendously, it also makes it easy to suffer from a serious problem of spam. Spammers are taking advantage of free hosting of sponsored advertisements and hyperlinks. Spam is prevalent in the blogosphere in different forms, as described in the following section.

1.1 Spam Types in Blogosphere

1.1.1 Splogs

Splogs are spam blogs where the post itself is used to promote a product or a service. It is also used to entice users familiar with the service to exploit search-engine reputation of the hosted service; to attract traffic from “neighboring”
blogs, etc. Additionally, free hosting services are the primary target for splogs due to the minimal cost of establishing one. A study reported at Google’s official Enterprise blog (Google Enterprise, 2009) shows that overall spam volume growth during the first quarter of 2009 is the strongest since early 2008, increasing an average of 1.2% per day. Across the blogosphere at large, a study in 2007 found that 56% of blogs which sent update notifications to the weblogs.com2 ping server were splogs (Kolari, 2007).

1.1.2 Comment Spam

Such spams are posted to a blog in no relation to the blog topic for the sole purpose of promoting service or a site. Unlike splogs, comment spam has been targeted to all types of blogs which allow commenting. The popular Akismet blog spam is a classification service, which runs hundreds of tests on the comment received and Akismet classifies 82% of submitted messages as spam correctly (Akismet, n.d.).

1.1.3 Trackback Spam

This category of spam takes advantage of the trackback ping feature of popular blogs to get links from them. While the HTML internals of comment submission forms may be changed to confuse spam robots without affecting legitimate users, trackbacks are transmitted by an HTTP-based protocol with a fixed API. The trackback specification makes no mention of verification, allowing spammers to inject arbitrary URLs into a trackback ping message along with camouflaging text of the spammer’s choosing. This has led to an abundance of trackback spam targeted at supporting blog software.

In this work, we only deal with short text-type spam just to narrow down the scope of our study. Spam comments are prevalent in blogs, WIKIs and all other online media where readers are allowed to freely post their comments. Blog comment spam is the act of posting comments to a publicly available blog which have nothing to do with the blog or post being commented on, and which are designed to direct users to other non-associated websites. There are several groups on the internet using blog spam in order to make money with advertising, click-throughs, adware installations and malware infections for stealing information (Valsmith, 2008). Comment spam generally contains hyperlinks to the spammers’ websites. These messages not only annoy web users, but also pollute web pages and waste Internet bandwidth. Another attractive incentive for spammers to create spam comments is the huge number of (hundreds of millions) web searches being conducted every day. Thus unethical content providers would like do anything that is necessary to make their contents highly ranked than they really should be in their market place.

The open interactive environment that blogs provides makes it a target for misuse or abuse, the major being the unavailability of a filter to decipher whether a comment is a spam or not. Spammers avoid being blacklisted or being tracked by registering, generating random or using dynamic IP address. CAPTCHAs (Complete Automated Public Turing test to tell Computers and Humans Apart) have remained popular, and successful to some extent in spotting spammers. However, naively designed CAPTCHAs is solved by machine learning algorithms, and overly complicated CAPTCHAs place too much burden on visitors, and discourage them to write any comment. Other popular approach, Keyword filtering can be effective, however, this approach requires manual monitoring of a list of locked keywords, and it sometimes filters out legitimate comments as well.

1.2 Comment Spam

Versus Email Spam

Email and comment spams share some common characteristics as well as a number of differences. While email spams are used solely to entice the reader to go to a site and/or to buy a
Related Content

Privacy-Enhancing Technologies
[www.igi-global.com/chapter/privacy-enhancing-technologies/21343?camid=4v1a](www.igi-global.com/chapter/privacy-enhancing-technologies/21343?camid=4v1a)

Probabilistic Inference Channel Detection and Restriction Applied to Patients’ Privacy Assurance
[www.igi-global.com/chapter/probabilistic-inference-channel-detection-restriction/62728?camid=4v1a](www.igi-global.com/chapter/probabilistic-inference-channel-detection-restriction/62728?camid=4v1a)
Ethical Concerns in Computer Science Projects
[www.igi-global.com/chapter/ethical-concerns-computer-science-projects/13474?camid=4v1a](www.igi-global.com/chapter/ethical-concerns-computer-science-projects/13474?camid=4v1a)

Formal Verification of Secrecy, Coercion Resistance and Verifiability Properties for a Remote Electronic Voting Protocol