Toward Spam 2.0: An Evaluation of Web 2.0 Anti-Spam Methods

Pedram Hayati, Vidyasagar Potdar

Digital Ecosystem and Business Intelligence (DEBI) Institute, Curtin University of Technology, Perth, WA, Australia
pedram.hayati@postgrad.curtin.edu.au, v.potdar@cbs.curtin.edu.au

***Abstract*—Spammers have proven very powerfully adaptable, if we thwart all current spam methods, they will find new loophole to use them. Blogs, comments, forums, opinions, online communities, wikis and tags are nowadays targets for their campaigns. This paper presents analysis of current anti-spam methods in Web 2.0 for spam detection and prevention against our proposed evaluation framework. The framework is a comprehensive framework to evaluate anti-spam methods from different perspectives. Our framework shows that the need for more robust methods which are prevention based, unsupervised and do not increase user and system interaction complexity is highly demanded.**

***Index Terms*—Spam, Web spam, Security, Anti-spam.**

I. Introduction

Unsolicited, anonymous, commercial and mass email messages, called *spam* is viewed as a serious problem for Internet, content quality and trust [1]. But email spam is not the only campaign for spammers as they always find new targets to achieve their desires. Web spam is defined as webpages are deliberately created to trick search engines into offering unsolicited, redundant and misleading search result as well as mislead users to unwanted and unsolicited webpage. Web spam is a recent domain that has been exploited by spammers. [2]. Besides from creating simple spam webpage or website [3], spammers nowadays, post promotional comments on blogs, write advertisement reviews for products, reply online forums threads with junk content, create eye-catching user profiles on online community websites, manipulate Wiki pages, and create misleading tags for their documents. These domains are example of Web 2.0 applications that relay on user-generated content, making them dynamic for both legitimate and spam content. The consequences of web *spamming* involve:

- Tricking search engine to rank spam and junk contents higher[2]. Hence it decreases quality of search engine results.
- Misleading users to view unsolicited content. For example as illustrated in Figure 1, spammer posted an attractive comment on user's blog along with a URL. However, The URL links to a kind of *linkfarm* page (Figure 2).

A report carried out by Sophos in 2008 revealed that every 3 seconds new spam-related webpage is created [4]. In other word, 23'300 spam-related webpages are created every day. These reports highlight an alarming point on the web and therefore research in spam prevention and detection is of prime importance to maintain quality of web content.

In this paper, we focus on evaluating current anti-spam methods for preventing and detecting spam content in blog, online forums, wikis, tags and online communities and we referred to as Web 2.0 anti-spam methods. We classify each anti-spam methods based on in its application and evaluate them along specific criteria. The perspectives presented here is to show to the best of our knowledge how much work has been done in each the spam domain and to highlight which domains require further investigation.

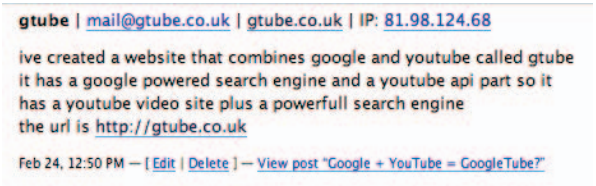


Fig.1 A sample spam comment

Fig.2 A sample linkfarm page

The rest of the paper is organized as follows. Section 2 outlines notion of spam in Web 2.0. Our proposed framework for evaluating current Web 2.0 spam methods is explained in Section 3. Section 4 studies anti-spam methods in each domain in detail and evaluates them against our proposed framework. We conclude our paper in Section 5 along with future research direction.

II. Web 2.0 Spam

Web 2.0 or the second generation of World Wide Web (WWW) is a new generation of webpages, moving from static webpages to dynamic and sharable content [5]. Websites such as Wikipedia, BlogSpot, del.icio.us, CNet, Facebook are examples of Web 2.0 websites. As mentioned earlier, Web 2.0 provides an environment where it is easy to

978-1-4244-3760-3/09/$25.00 ©

generate both legitimate and spam content. Hence spammers have created new campaigns Web 2.0 websites. In this paper we refer spam content in Web 2.0 as *Web 2.0 spam* or simply *Spam 2.0*. Although there is research in the realm of Web 2.0 spam, we believe that existing literature has been limited by the lack of real world examples of spam in each of the Web 2.0 domains. Hence, the rest of this section we discuss methods that are used by spammers to distribute spam 2.0.

A. Hosting blogs, writing blog comments and making trackbacks

Spammers create fake blogs (called *splog*) for promoting junk/hijack content or generating link farms [6]. The content inside a splog can be used for misleading search engines ranking algorithms or misleading users to an unwanted website. Factors such as easy generation of blogs, the numbers of free blog hosting services, and rapid indexing of blog contents by search engines make blog as an attractive platform for spammer [7]. A sample splog is presented in Figure 3.

Comment in the blog post is a feature can be used by blog visitors to share their opinion with blog's author. This feature can be utilized by spammer to distribute promotional, fake and junk content which is called *comment spam*. Spammers can employ comment spam to place links from a legitimate blog to their spam websites to mislead search engine algorithms and users. Figure 1 presents an example of comment spam.

Trackback is a method for notifying author (e.g.: blog author) when somebody has linked to one of their posted documents [8]. This notification is usually in the form of comment at the end of original blog post which has link back to cited document. Trackback is misused by spammers place a link to their campaigns from legitimate domains.

Fig.3 A sample splog on BlogSpot

B. Posting new threads in online forums

Online forums are type of Web 2.0 applications for holding discussion and comments of users. Forums can also be utilized by spammers to distribute spam content. By registering a username in forums and posting either new discussion or replying to other discussions spammers distribute their junk content. This type of spam domain is *forum spam*. Figure 4 presents an example of forum spam. Spammers created a thread inside a forum contain keywords and URLs to their campaigns. Keywords in this example are used to give the wrong impression on actual content of linked website to the search engine crawler.

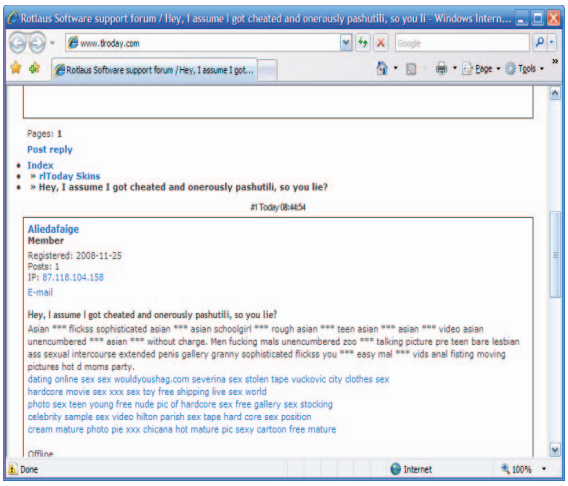
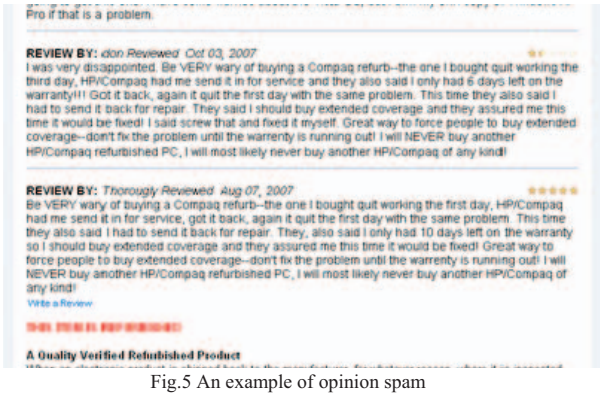


Fig.4 A sample forum spam

C. Writing reviews

Reviews are very valuable resources for both customers and manufactures. Customers can read reviews to find others opinions about products, so they can decide better on buying products. Manufactures, on the other hand, can use reviews to get valuable feedbacks for their products and optimize their product in order to satisfying customers' requirements. Opinion or review spam refers to advertising, promoting or misleading reviews on products in merchant websites [9]. Opinion spam is used to promote a product or giving unfair review on products or services. Additionally, some opinion spams contain advertisement and irrelevant content [10]. Figure 5 illustrate an example of two opinion spam with more similar content with the aim to damage reputation of a one brand name.


Pro if that is a problem.

REVIEW BY: *don Reviewed Oct 03, 2007*
I was very disappointed. Be VERY wary of buying a Compaq refurb--the one I bought quit working the third day, HP/Compaq had me send it in for service and they also said I only had 6 days left on the warranty!!! Got it back, again it quit the first day with the same problem. This time they also said I had to send it back for repair. They said I should buy extended coverage and they assured me this time it would be fixed! I said screw that and fixed it myself. Great way to force people to buy extended coverage--don't fix the problem until the warrenty is running out! I will NEVER buy another HP/Compaq refurbished PC, I will most likely never buy another HP/Compaq of any kind!

REVIEW BY: *Thoroughly Reviewed Aug 07, 2007*
Be VERY wary of buying a Compaq refurb--the one I bought quit working the first day, HP/Compaq had me send it in for service, got it back, again it quit the first day with the same problem. This time they also said I had to send it back for repair. They, also said I only had 10 days left on the warranty so I should buy extended coverage and they assured me this time it would be fixed! Great way to force people to buy extended coverage--don't fix the problem until the warrenty is running out! I will NEVER buy another HP/Compaq refurbished PC, I will most likely never buy another HP/Compaq of any kind!
Write a Review

A Quality Verified Refurbished Product
Fig.5 An example of opinion spam

D. Creating user profiles for online social activities

Online communities or social networking websites such as Facebook and MySpace provide a platform for participating in many social activities (e.g.: making and find new

friends, sharing photos/videos, chatting etc). Each user represented by a *profile* which can be virtual representation of a user. Unfortunately, Spammers are not absent from these online platforms. They create deceptive profiles which can be used to mislead users to unsolicited webpages, spreading fake information and distributing malwares [11]. This type of spam refers as *social spam*. Additionally, recently there has been some reports on existence of spam in video sharing website such as YouTube [12]. Spammers put spam response to some legitimate videos in order to attract genuine users to their entries.

E. Modifying Wiki pages

Wiki spam is very difficult type of spam to detect and prevent. Spammer modifies wiki pages in order to back link to their targets or injects false content (e.g.: fake references) inside Wiki pages. Additionally spammer misuses Wiki's features (e.g.: Wiki's ignore tag) to inject HTML codes or link inside Wiki pages which are hidden [13]. Currently this kind of spam is detected and removed manually by Wiki's editors. Figure 6 illustrates an example of Wiki spam that spammer has modified URL address to their target.

Fig.7 An example of Wiki spam

F. Making tags

Tags are descriptive string for annotating shared resources such as Blog posts. User can choose tags for their documents in order to highlight main related keywords. By making false and not-related tags, spammer tries to attract more viewers or boost visibility of their resources. Figure 8 shows an example of tag spam in video sharing websites.

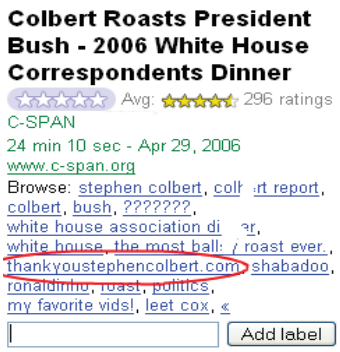


Fig.8 A sample tag spam.

In the next section we propose our evaluation framework for analyzing Web 2.0 anti-spam methods. The evaluation framework is used in Section IV to evaluate Web 2.0 anti-spam methods in each domain.

III. Proposed Framework

In order to tackle spam concern in Web 2.0, there have been many works proposed in the literature. In this section we propose our evaluation framework for analyzing anti-spam methods based on different criteria. The framework gives comprehensive view on each anti-spam method, i.e.: what are drawbacks/advantages of each method, in which domain they can/cannot be applied etc.

Our evaluation framework consists from 2 main criteria and 8 different sub-criteria as follows:

1. Is the method a detection strategy?
   a. Is it a language dependent method?
   b. Is the method content based or meta-data based?
   c. Does the method use supervised, semi-supervised or non-supervised machine learning approach?
   d. Is the method behaviour based?
   e. Does the method decrease user-interaction convenient?
2. Is the method a prevention strategy?
   a. Does the method prevent spammer to use user network resources?
   b. Does the method increase complexity of user-interaction with system?

Anti-spam methods can be categorized into category or strategy – detection and prevention [3, 14]. In detection strategy, anti-spam method looks inside content and search for spam patterns. For example if the title of new incoming email contains "$$$$ BIG MONEY $$$$" it is an example of spam pattern and would be marked as spam. On the other hand, prevention strategy based methods stop spammers to enter into the system. It tries to spot spammers before they enter to the system or distribute their spam content. Example of these methods can be blacklisting and Whitelisting of IP addresses. Hence main 2 criteria differentiate methods based on detection and prevention strategy. Prevention based methods put time and computational cost on client side but they are not wasting server resources on the other hand detection methods consume server-side resources but they are not increasing user and system interaction convenience.

Some of detection methods are based on English language and English grammar so they are language specific and can not be applied to detect non-English language spam. So in question 1.a methods are investigated whether or not they can be applied on non-English language spam content.

Question 1.b is about different places methods are looking for to find spam pattern. Some methods look inside actual content such as email body message while others look inside meta-data such as email headers. Comparatively, Content based methods require more time and process to

classify the content. Meta-data based methods are reasonably faster in detection however the amount of features for classifying task is limited.

Supervised, semi-supervised, or unsupervised methods are focus of question 1.c. Supervised and semi-supervised methods require manual efforts for labelling the training datasets and they need frequent updates for their training datasets to be able to detect new spam contents. In other way, supervised and semi-supervised methods are one step behind spammers and they must be regularly updated. These methods put computational cost on server-side. Some methods track user behaviour inside software hence they create a *profile* for each user inside system. Question 1.d addresses these methods. Behaviour based detection methods are dependable on historical data. Spammers have shown that they changed their behaviour and they attempt to imitate legitimate user behaviour. Hence these methods are not able to detect them sufficiently. Additionally for studying behaviours methods require a couple of time and storage to create profiles which consume server storage space and time.

Finally in Question 1.e, some spam methods put a lot of pressure on spam detection but on the other hand decrease user convenient on interacting with the system. In this question we try to study this aspect of method that how method deals with user convenient.

In the other main criteria, the focus of framework is on prevention based methods. Prevention based methods have advantages to put computational and time cost on client side; However, they may/can increase complexity of system. Question 2.a investigates whether anti-spam method stops spammer entering into the system or not. And Question 2.b as mentioned earlier checks whether or not method decreases users' convenience.

Based on this framework in the next section we investigate recent anti-spam methods in Web 2.0. We discuss about their advantage and drawbacks as well as their gaps.

IV. Evaluation

In this section we investigate recent Web 2.0 anti-spam methods against our proposed evaluation framework. We divide anti-spam methods into 3 categories i.e. blog (include comment and trackback), review, social spam. In each category we first give brief introduction to each method then we evaluate method against our framework. We begin by evaluation of blog, comment and track back spam.

A. Blog, comment and trackback spam

Research in blog spam is relatively in its infancy. One of the first articles to talk about blog spam was presented in early 2004 [15] which has just limited to existence of spam in blog.

Mishne [16] presented one of the first method to detect comments spam in blogs using language model disagreement. This method compare language of blog post, comment, and the webpage linked in comment then by comparing them it can classify comment as spam or legitimate comment. The main advantage of this method is that it is unsupervised method. Hence it doesn't need any training dataset. We believe that this detection method can be applied on non-English language contents as well since it deals with comparison of words inside the blog, comment and linked webpage so it does not weed into grammar of particular language. This method does not put any effort for user interaction with the system; hence it does not decrease user convenience. Additionally this method is content based method.

In [17] authors proposed a collaboration spam detection method for detecting link spam inside comment and trackback. The idea of this method is to manually identification of spam by genuine users and to share them among other users. This method can be applied for non-English language content since the detection phase is independent from the content. It is kind of supervised method which increases complexity of user interaction with the system since user needs to classify legitimate comment from spam comment.

Authors in [18] proposed an idea to detect blog spam based on vocabulary size strings inside blog post, comment and trackback. The proposed method extract frequencies of sub strings in all blog posts and if it is higher than specific threshold it marks blog as splog. Their method can be applied on non-English language content as authors themselves mentioned that they found Japanese language splog by using their method and it independent from language grammar and structure. This detection method looks inside blog posts content and is a semi-supervised method since sometime they need to differentiate between spam strings and post templates. This method does not increase complexity of system.

Methods presented in [6, 19-21] are using supervised machine learning approach to detect splogs. They extract some features from each blog (e.g.: Bag-of-word, URL segmentation, Update pings, Time of post etc) compare features against training dataset by using Support Vector Machine algorithm (SVM). These detection methods can be applied on non-English language content if extracted features are not depended on grammar (such as link structure that is language independent). They are both content and meta-data based since some feature are extracted from content (e.g.: bag-of-word) or/and some from meta-data (e.g.: tokenized URLs). These methods do not have any influence on user and system interaction.

In the recent work [22] authors used writing behaviour (include: writing interval, writing structure, writing topic) in order to detect splog from legitimate blog. The main assumption of their work is that spammer tries to focus on same topic with the same amount of content compare with legitimate bloggers with various topic and different blog post length. Their method does not depend on specific language structure hence it can be applied on non-English contents. It is a content-based method since it looks into content of blog post to find writing behaviours. Authors used SVM, Naïve Bayesian and C4.5 machine learning algorithms in their method hence it is supervised method. Additionally since they are looking for writing behaviour in blog

posts this method is behaviour based anti-spam detection. There is no additional task need to done by user for spam detection hence this method does not increase user and system interaction convenience.

In [23] splog detection task is done by comparing rate of copied content. Splogs usually are made up of plagiarism content from other sources. Authors use this feature to classify between splog and legitimate blog. The content of blog post compare with other blog post and if the copy rate be higher than specific threshold, content would be marked as spam. Authors believed that their method can be applied on non-English as well as English content. Method is a kind of content based detection method and does not increase user-and-system-interaction complexity. The main advantage of this work is that it is an unsupervised spam detection method.

Overall, according to our framework all the blog, comment, and trackback anti-spam methods are detection based and developing prevention based anti-spam methods in this domain are interested.

B. Review/Opinion Spam

Research in developing anti-spam method to detect or prevent spam content in reviews is quite young. The work presented by [10] uses 36 features to do the classification task. The presented detection method employs supervised machine learning algorithm (logistic regression). Depend on feature it is language independent and looking inside the content to find spam patterns.

This spam domain is unique and the need for more robust anti-spam methods is highly demanded.

C. Social Spam

The method presented by [24] is a spam detection method 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 can/cannot be language independent. There is no pressure on user side for differentiation among genuine users an spammers.

In [12] authors propose 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.

As mentioned earlier, tags are descriptive string for annotating shared resources such as Blog posts. As they become more popular they are at risk to be target for spammers. The authors in [25] proposed and idea of tagging system which can be robust in front of spamming techniques, this tagging system counts number of coincident (or common) tags amongst other users and assign document a relevance ranking number. By looking at ranking number it can differentiate among spam and legitimate tags. This method is language independent and content based.

This domain of spam battle is young and not many works has been done so far. Additionally, current solutions are not effective – the method presented in [24] is just 30-40% better than random spam detection. All the methods are supervised detection based. Since Web 2.0 is collaborative environment and is evolving/changing everyday, the need for prevention based and unsupervised methods in this environment are highly demanded.

According to our survey, apparently there is no specific work in other domains such as Wiki spam and forum spam. Specially, wiki spam nowadays is hard to detect and need manual efforts. Additionally these domains are target of more spammers since forums and wiki pages are indexed more often by search engines and more trusted by users.

V. CONCLUSION

The framework that proposed in this paper gives a comprehensive viewpoint on current literature of spam combating and it reveals drawback of current literature methods. It also shows that in which domains there are a few works and how they deal with user and system interaction complexity. We found that most of the proposed anti-spam methods use server-side resources. So not only spammer but also these detection-based methods consume server-side resources. On the other side, prevention based methods put componential cost on the user side. However they increase complexity of system for user-interaction. Most of the current methods are supervised and need up-to-date datasets which in current growing rate of Web 2.0 application is a resource consuming task. On some domains such as Wiki and forum there is no particular method. So the need for more robust methods which are prevention based, unsupervised and do not increase user and system interaction complexity is highly demanded.

VI. REFERENCES

[1] A. Cournane and R. Hunt, "An analysis of the tools used for the generation and prevention of spam," *Computers & Security,* vol. 23, pp. 154-166, 2004.

[2] Z. Gyongyi and H. Garcia-Molina, "Web spam taxonomy," in *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web*, Chiba, Japan, 2005.

[3] H. Paul, K. Georgia, and G.-M. Hector, "Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges," *IEEE Internet Computing,* vol. 11, pp. 36-45, 2007.

[4] net-security.org, "Latest spam statistics," in *http://www.net-security.org/secworld.php?id=6056*, 2008.

[5] T. O'Reilly, "What Is Web 2.0," in *http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html*: O'Reilly Network, 2005.

[6] L. Yu-Ru, S. Hari, C. Yun, T. Junichi, and L. T. Belle, "Splog detection using self-similarity analysis on blog temporal dynamics," in *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web* Banff, Alberta, Canada: ACM, 2007.

[7] P. Kolari, A. Java, and T. Finin, "Characterizing the Splogosphere," in *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem*, 2006.

[8] M. Kimura, K. Saito, K. Kazama, and S.-y. Sato, "Detecting Search Engine Spam from a Trackback Network in Blogspace," in *Knowledge-Based Intelligent Information and Engineering Systems*, 2005, pp. 723-729.

[9] N. Jindal and L. Bing, "Analyzing and Detecting Review Spam," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Con-*

ference on, 2007, pp. 547-552.
[10] J. Nitin and L. Bing, "Opinion spam and analysis," in *Proceedings of the international conference on Web search and web data mining* Palo Alto, California, USA: ACM, 2008.
[11] S. Webb, J. Caverlee, and C. Pu, "Social Honeypots: Making Friends with a Spammer Near You," in *Proceedings of the Fifth Conference on Email and Anti-Spam (CEAS 2008)*, Mountain View, CA, 2008.
[12] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, C. Zhang, and K. Ross, "Identifying Video Spammers in Online Social Networks," in *AIRWeb '08* Beijing, China, 2008.
[13] LinkSleeve, "Wiki Spam," in *http://www.linksleeve.org/wiki-spam.php*, 2008.
[14] P. Hayati and V. Potdar, "Evaluation of Spam Detection and Prevention Frameworks for Email and Image Spam - A State of Art," in *The 2nd International Workshop on Applications of Information Integration in Digital Ecosystems (AIIDE 2008)* Linz, Austria, 2008.
[15] P. McFedries, "Technically Speaking: Slicing the Ham from the Spam," *Spectrum, IEEE,* vol. 41, pp. 72-72, 2004.
[16] G. Mishne, D. Carmel, and R. Lempel, "Blocking Blog Spam with Language Model Disagreement," in *In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.
[17] S. Han, Y. Ahn, S. Moon, and H. Jeong, "Collaborative blog spam filtering using adaptive percolation search," in *InWWW 2006 Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.
[18] K. Narisawa, Y. Yamada, D. Ikeda, and M. Takeda, "Detecting blog spams using the vocabulary size of all substrings in their copies," in *WWE 2006 3rd Annual Workshop on the Weblogging Ecosystem* Edinburgh, Scotland, 2006.
[19] P. Kolari, T. Finin, and A. Joshi, "SVMs for the blogosphere: Blog identification and splog detection," in *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, Stanford University, California, 2006.
[20] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi, "Detecting Spam Blogs: A Machine Learning Approach," in *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)*, Stanford University, California, 2006.
[21] D. Sculley and M. W. Gabriel, "Relaxed online SVMs for spam filtering," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* Amsterdam, The Netherlands: ACM, 2007.
[22] W. Liu, S. Tan, H. Xu, and L. Wang, "Splog Filtering based on Writing Consistency," in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2008.
[23] T. Takeda and A. Takasu, "A splog filtering method based on string copy detection," in *Applications of Digital Information and Web Technologies, 2008. ICADIWT 2008. First International Conference on the*, 2008, pp. 543-548.
[24] A. Zinman and J. Donath, "Is Britney Spears spam," in *Fourth Conference on Email and Anti-Spam* Mountain View, California, 2007.
[25] K. Georgia, E. Frans Adjie, Zolt, G. n, ngyi, H. Paul, and G.-M. Hector, "Combating spam in tagging systems," in *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web* Banff, Alberta, Canada: ACM, 2007.