SocialSpamGuard: A Data Mining-Based Spam Detection System for Social Media Networks

Xin Jin
Department of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Ave.
Urbana, IL USA
xinjin3@illinois.edu

Cindy Xide Lin
Department of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Ave.
Urbana, IL USA
xidelin2@illinois.edu

Jiebo Luo
Kodak Research Laboratories
Eastman Kodak Company
1999 Lake Avenue
Rochester, USA
jiebo.luo@kodak.com

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Ave.
Urbana, IL USA
hanj@cs.uiuc.edu

ABSTRACT

We have entered the era of social media networks represented by Facebook, Twitter, YouTube and Flickr. Internet users now spend more time on social networks than search engines. Business entities or public figures set up social networking pages to enhance direct interactions with online users. Social media systems heavily depend on users for content contribution and sharing. Information is spread across social networks quickly and effectively. However, at the same time social media networks become susceptible to different types of unwanted and malicious spammer or hacker actions. There is a crucial need in the society and industry for security solution in social media. In this demo, we propose *SocialSpamGuard*, a scalable and online social media spam detection system based on data mining for social network security. We employ our GAD clustering algorithm for large scale clustering and integrate it with the designed active learning algorithm to deal with the scalability and real-time detection challenges.

1. INTRODUCTION

We have entered the era of social media networks, e.g., Facebook, Twitter, YouTube and Flickr. Internet users now spend more time on social networks than search engine. Business entities or public figures set up social network pages to enhance direct interaction with online users.

Take Facebook Page as an example, in addition to 500 million users, there are over 14 million (the number is keep growing) pages from various categories, such as company, product/service, musician/band, local business, politician, actor/director, artist, athlete, author, book, health/beauty, movie, cars, clothing, community, food/beverages, games, toys, government organization, interest, sports, TV channel, TV show, and website. The current most popular page (Texas Hold'em Poker) has 40 million fans, and there are around 3000 pages that have more than 500,000 fans, with an average of 2 millions. Fans not only can see information submitted by the page, but also can post comments, photos and videos to the page.

Social media websites allow users freely distribute and share information to friends. Information can spread very fast and easily within the social media networks. Because of this, such websites expose to various types of unwanted and malicious spammer or hacker actions. There is a crucial need in the society and industry for a security solution in social media. Social media websites need to be clean for long term success. A company/brand page on social media also needs to be clean to reduce the risk of damaging its reputation. Virus links from the spams could lead to personal or business loss and damage.

There have been some studies on detecting spam emails [8, 15], spam messages [12], spam images [2], spam video [1], web spam [13], spammers [11] [9] [14], etc. However, one of the major challenges of spam detection in social media is that the spams are usually in the form of photos and text, and in the context of large scale dynamic social network. We need a comprehensive solution which can consider text, photos and the social network features, and also be scalable and capable of performing real-time detection.

In this demo, we propose propose *SocialSpamGuard*, a scalable and online social media spam detection system based on data mining for social network security. The major advantages of the proposed approach can be summarized as follows:

1. Automatically harvesting spam activities in social network by monitoring social sensors with popular user bases;

2. Introducing both image and text content features and social network features to indicate spam activities;

Proceedings of the VLDB Endowment, Vol. 4, No. 12

3. Integrating with our GAD clustering algorithm to handle large scale data;

4. Introducing a scalable active learning approach to identify existing spams with limited human efforts, and perform online active learning to detect spams in real-time.

2. SOCIAL MEDIA NETWORK MODEL

As shown in Figure 1, we model a typical social media network as a *time-stamped heterogeneous information network* $G = \langle V, E \rangle$. V is the set of different types of nodes, such as users (U), pages (P) and posts (Q) (including text description and/or images/videos (I), with the time stamp). E denotes the set of links between nodes, for example, friendship/following links between users, fan/favorite links between users and pages. Images are indirectly-linked together by content similarity (dashed lines).

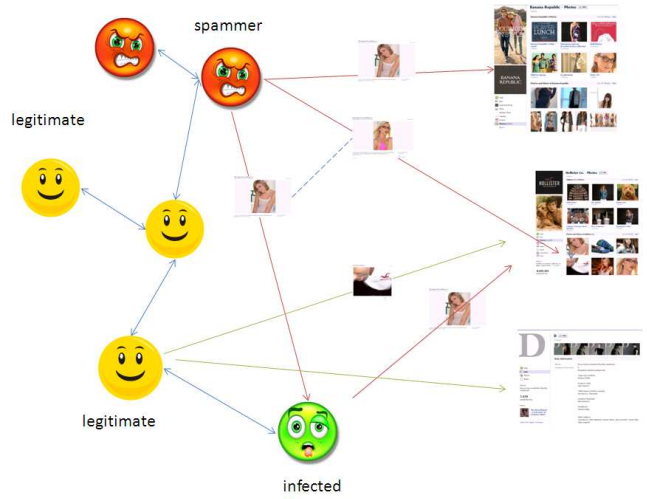

Figure 1: Heterogeneous Information Network for Social Media. A red face is a spammer, a yellow smile face is a legitimate user, a yellow face turned to green color is an infected user. The blue directed line is the friendship/following link. A red arrow is a spam post, while a green arrow is a ham post.

Posting is one of the predominant user activities in social media. We spend most of our time in social media, such as Facebook/Twitter, on posting or checking the posts of friends or favorite pages.

The posts can be generally labeled as two categories: **spam** (unwanted, irrelevant, promotional or harmful social posts) and **ham** (legitimate social posts). There are three types of users: spammer, legitimate user and infected user. Infected user are legitimate user who send spams after being infected by virus. Our **goal** is to identify the spam posts sent from spammers and infected users.

3. SYSTEM FRAMEWORK

As shown in Figure 2, the system architecture works as follows. In the first stage, we collect historical social media data, extract both content (including text and images) and social network features, perform active learning to build classification model and identify spams. In the second stage, we monitor the real-time activity of the social network and perform online active learning, make prediction and send alarms to clients about detected spams, collect feedbacks from clients and update the model.

3.1 Feature Extraction

In this section, we extract the image content features, text features and social network features to describe the posts in social media network.

3.1.1 *Image Content Features*

We extract image content features [7] [4] [5], such as color histogram, color correlogram, CEDD [3], Gabor features, edge histogram and SIFT [10], to help build classifier to identify spam photos. Many spam photos are beauty, high quality and attractive, in order to attract users to click on it.

In addition, based on our observation, many spam photos look different from the legitimate photos of a page. For example, in the Hollister Co. Facebook page in Figure 3, a photo of bag is not related to any products of Hollister, so it will probably be a spam. To compute image similarity in the social network, we use the SimLearn algorithm [7].

3.1.2 *Text Features*

Text features are extracted from image-associated content, such as caption, description, comments, and URLs. We expect legitimate images to avoid sensitive words and have enough comments and "normal" URLs. We list several text-related features as follows:

- The ratio of content which consists of non-English words.
- The number of sensitive words, the number of comments/likes.
- The reputation of comment authors.
- Whether the contained URL is a short URL (or duplicate short URLs) which leads to spam website (e.g., both 'http://nxy.in/xxhp1' and 'http://nxy.in/3tw2s' point to the same spam website 'xxxblackbook.com').

3.1.3 *Social Network Features*

Considering the social network information, we extract the third set of attributes which consists of individual characteristics of user profiles and their behaviors in the network. Both legitimate users and spammers have certain kinds of patterns.

Legitimate users have many legitimate friends, while spammers almost never reply to comments and many spammers are registered as beautiful females. A spammer may have several photos of herself, or use celebrity's name and photos. A spammer tends to post to popular pages (mostly over 500,000 fans) to gain high chance of exposure.

The spammer attracts other users to add himself/herself as a friend, instead of initiating the friend requests to reduce the risk of being detected as spammer, because spammers have lower friend request acceptance rate than legitimate users and systems like Facebook treat a user with low friend request rate as potential spammer.

Many spammers, with the help of computer program or the spammer is just a computer program instead of being a real person, sometimes post the same photo or several

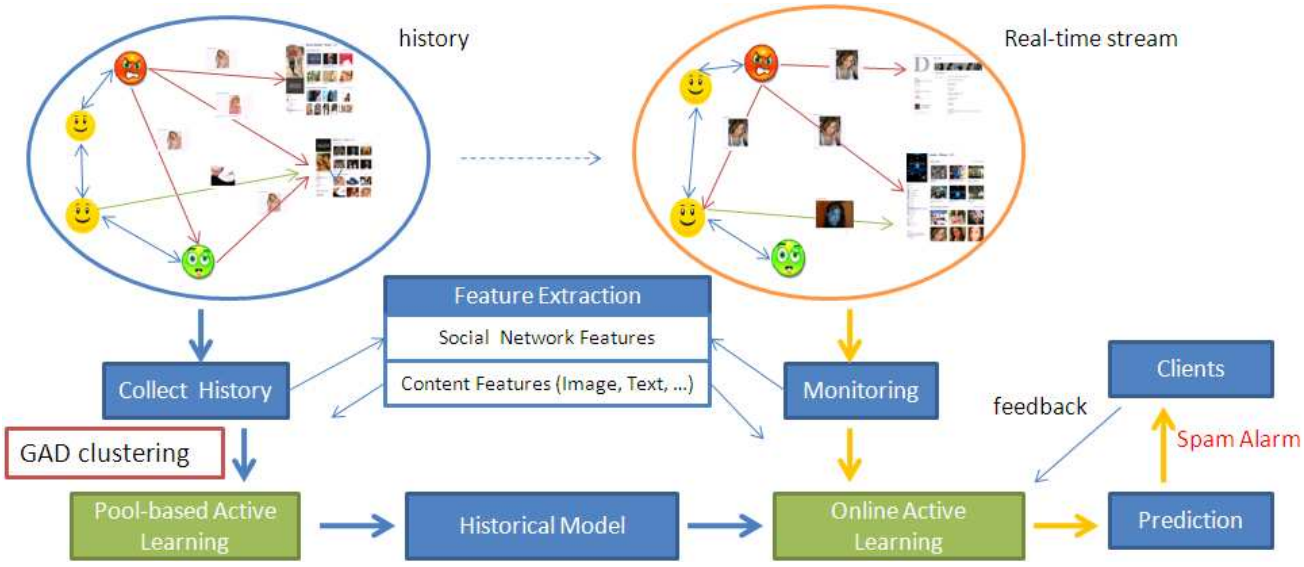

Figure 2: System Architecture.

similar photos to many popular pages or the friends during a short time range.

We consider all of these observations as features which are unique in the scenario of a social network.

3.2 Scalable Active Learning for Historical Data

Due to the huge amount of posts (over billions) on social media, manually checking every post to pick up the spams is impossible. We propose the following scalable active learning approach to manually verify as many spams as possible and as few hams as possible.

(1) Generate an initial set of instances for labeling and build initial classifier.

(2) Prediction and ranking of remaining unlabeled instances (which is a huge number) using the existing classifier. Sort the test posts in decreasing order according to the ranking score and divide them into blocks.

(3) Obtain an additional set of labeled posts. Such set is formed by examining the top blocks in both order. Uncertain posts and a random set are also included.

(4) Add the new labeled set to the training pool, and update the classification model.

(5) Iterate steps 2 to 5 until satisfying a stop criteria, such as the maximum number of iterations or the minimum number of additional spams detected.

GAD Clustering for Smarter Sampling. Because of the huge number of posts, randomly sampling may not be a good choice due to the uneven distribution and duplicate (or near duplicate) posts. To generate a smarter sample in the active learning procedure, we use our algorithm GAD [6] to perform large scale clustering of the posts into large number of clusters and make sampling from the clusters to increase diversity and avoid duplicates.

3.3 Online Active Learning for Real-Time Monitoring

After building the classification model based on the historical data, we can start real-time monitoring of new posting activities in the social media network. For each new instance s, we first make prediction based on the trained model, if it is uncertain, send the instance for human labeling and add s to pool T_{new}. If $|T_{new}|$ becomes bigger than a threshold, add it to the training pool for model retraining.

4. CASE STUDY

We showcase the prototype system using Facebook as the example application. We choose popular pages with over 500,000 fans (an average of 2 million fans for each page) as basic sensors to monitor the public posting activities in the social network.

As an example, Figure 3 shows the Hollister Co. page on Facebook[1]. Hollister is popular American lifestyle clothing brand targeting on young people. As of March 28, 2011, its Facebook page has 4,608,404 fans and 5,100 user added photos/videos, most of which are photos. In the figure, the section marked as red rectangle lists the top 6 recently added photos, 4 of which are detected as spams (marked as red X). Take the first one as an example, if we click on it, it shows the following description: "I am a very sweet woman and I am seeking for a gorgeous man to share a joy night with.See how gorgeous I am at http://nxy.in/xxhp1".

Note. This demo uses Facebook as the example since it is currently the most popular social media website. However, the technique proposed can be easily generalized and applied to other social sharing websites, such as Twitter and YouTube, to provide a more comprehensive social network security solution.

5. DEMONSTRATION

The design and development of the proposed system involve challenging issues in database, data mining and computer vision. We will thoroughly present our social media spam detection system in the demonstration.

We will present the technical details of the system, including the spam features, algorithms and efficient implementation. The rationale behind the design will be analyzed, espe-

[1]http://www.facebook.com/hollister?sk=photos

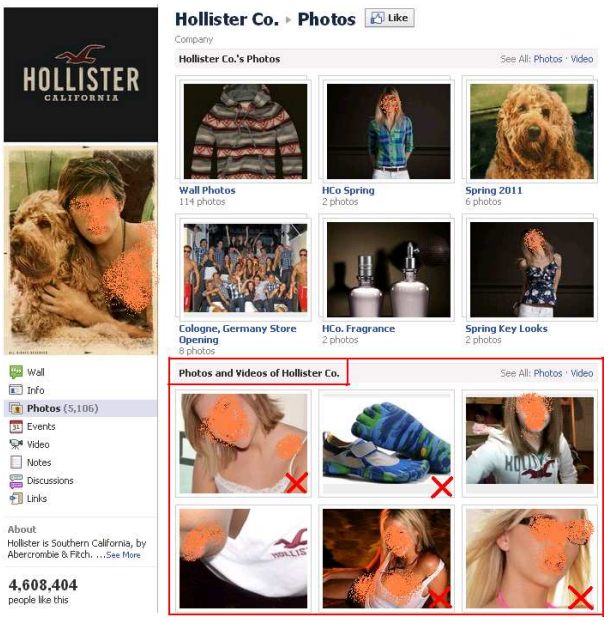

Figure 3: The Hollister Co. page on Facebook, accessed on March 28, 2011. The section "Photos and Videos of Hollister Co." (marked as red rectangle) lists the user added photos/videos in time decreasing order. Among the top 6 most recent photos, 4 of which are detected as spams (marked as red X). For privacy consideration, we have mosaicked the photos.

cially on the scalability and accuracy issues, in order to show how our system can handle the huge number of posts and monitor real-time social activities in social media to identify spams.

During the demonstration, we will setup a website to show the recently detected spam messages/photos and the corresponding spammers or infected users. In this way, the audience can obtain intuitive understanding about the essence of online social spam detection for social media.

Note that it is conceivable that many social websites may be using some of the spam filtering heuristics mentioned in this paper, though to the best of our knowledge these have not been well documented in publications before. In addition, we found that spams are widely present in the dataset we collected from Facebook, so the existing spam filtering process, if any, is not effective enough.

6. ACKNOWLEDGMENTS

This work was sponsored by Eastman Kodak Company and supported in part by NSF grant IIS-09-05215, MURI award FA9550-08-1-0265 and NS-CTA W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the sponsors. We thank the owners of the publicly accessible photos used in this paper.

7. REFERENCES

[1] F. Benevenuto, T. Rodrigues, V. Almeida, J. M. Almeida, C. Zhang, and K. W. Ross. Identifying video spammers in online social networks. In *AIRWeb*, pages 45–52, 2008.

[2] B. Byun, C.-H. Lee, S. Webb, and C. Pu. A discriminative classifier learning approach to image modeling and spam image identification. In *CEAS*, 2007.

[3] S. Chatzichristofis and Y. Boutalis. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. *Lecture Notes in Computer Science*, pages 312–322, 2008.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, April 2008.

[5] T. Deselaers, D. Keysers, and H. Ney. Features for image retrieval: an experimental comparison. *Information Retrieval*, 11(2):77–107, 2008.

[6] X. Jin, S. Kim, J. Han, L. Kao, and Z. Yin. A general framework for efficient clustering of large datasets based on activity detection. *Statistical Analysis and Data Mining*, 4(1):11–29, 2011.

[7] X. Jin, J. Luo, J. Yu, G. Wang, D. Joshi, and J. Han. iRIN: image retrieval in image-rich information networks. In M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, editors, *WWW*, pages 1261–1264. ACM, 2010.

[8] J. S. Kong, B. A. Rezaei, N. Sarshar, V. P. Roychowdhury, and P. O. Boykin. Collaborative spam filtering using e-mail networks. *IEEE Computer*, 39(8):67–73, 2006.

[9] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots + machine learning. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 435–442, 2010.

[10] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2*, page 1150, 1999.

[11] B. Markines, C. Cattuto, and F. Menczer. Social spam detection. In *AIRWeb*, pages 41–48, 2009.

[12] C. Shekar, S. Wakade, K. J. Liszka, and C.-C. Chan. Mining pharmaceutical spam from twitter. In *ISDA*, pages 813–817, 2010.

[13] S. Webb, J. Caverlee, and C. Pu. Introducing the webb spam corpus: Using email spam to identify web spam automatically. In *Proceeding of the Third Conference on Email and Anti-Spam (CEAS)*, 2006.

[14] S. Webb, J. Caverlee, and C. Pu. Social honeypots: Making friends with a spammer near you. In *Proceeding of the Fifth Conference on Email and Anti-Spam (CEAS)*, 2008.

[15] K. Yoshida, F. Adachi, T. Washio, H. Motoda, T. Homma, A. Nakashima, H. Fujikawa, and K. Yamazaki. Density-based spam detector. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 486–493, 2004.