SPAM PROOF TAGGING SYSTEM USING TRUST MODELING ALGORITHM

[1]SEEMA BHURAVANE, [2]DIPTI PATIL

[1,2]Computer Engineering Department PIIT New Panvel,
Mumbai University, India
E-mail: [1]skbhuravane@gmail.com, [2]dypatil75@gmail.com

Abstract— Tagging in online social networking site is very popular these days, as it facilitates search and retrieval of various resources such as text, images, videos, etc. Despite the advances in social networking over the past few decades, one of the important challenges that user continuously facing is spam. Noisy and spam annotations often make it difficult to perform an efficient search. The shared content is sometimes assigned with inappropriate tags for several reasons. Users may make mistakes while tagging and irrelevant tags and content may be maliciously added for their advertisement or self-promotion. Consequently, assigning tags to resources has a risk that wrong or irrelevant tags eventually prevent users from the benefits of annotated content. One important challenge in tagging is to identify the legitimate tags for given content, and at the same time, to eliminate spam tags. Trust can predict the future behavior of users to avoid undesirable influences of untrust-worthy users. Here we proposed a trust-worthy system that has been designed with the objective to minimize spam tagging and posting in social networking sites with the adaptation of classification algorithms.

Keywords— Tagging, Tagging system, Trust modeling, Tag spam.

I. INTRODUCTION

Social networking sites make it possible to users to form social relations among people who share similar interests, real life activities or connections. Whenever information is exchanged on the Internet, spammers are everywhere and they try to take advantage of the information exchange structure for their own benefit, while troubling and spamming others. Before social tagging became popular, spam content was observed in various domains. First in e-mail, and then in Web search networks have been also influenced by malicious peers, and thus various solutions based on trust and reputation have been proposed, which dealt with collecting information on peer behavior, scoring and ranking peers, and responding based on the scores.

Social tagging became popular with the launch of various sites like Delicious and Flickr. After that, different social tagging systems have been built to support tagging of a variety of resources like text, images. For given a particular resource, tagging is a process where a user assigns a tag to an object. Most tagging systems such as Delicious and Flickr are collaborative in nature in that they allow users to share and peruse tags and resources from other members of the community. On Delicious, a user can assign tags to a particular bookmarked URL. On Flickr, users can tag photos uploaded by them or by others. Whereas Delicious allows each user to have her personal set of tags per URL, Flickr has a single set of tags for any photo. On blogging sites like Blogger, Wordpress, Livejournal, blog authors can add tags to their posts. On micro-blogging sites like Twitter, hash tags are used within the tweet text itself. On social networking sites like Facebook users often annotate parts of the photos. Users can also provide tagging information in other forms like marking something as "Like" on Facebook. Upcoming event sites can allow users to comment on and tag events. Despite the advances in social networking sites over the past few decades, one of the important challenges that user continuously facing is spam. Malicious users continue to innovate ways to take advantage of public trust. Literature survey shows that the spam on Facebook and YouTube websites are extremely higher than what may be noticed on other social media websites.

Tagging services in social networks, e.g., Flickr, Delicious, YouTube, have grown in significance on the Internet based on the number of participating users. In a typical tagging system, each specific resource like post, photo, URL is annotated with some tags. Resource annotators are the users, who have annotated a specific resource with some tags and the relation <tag, resource> that annotates a resource with a tag is called an annotation. Annotation preserves the association between the tag and resource. When the user issues a tag in search bar, the system retrieves resources associated with this tag. Then, the user may collect some of the resources, and annotates it with some tags. By literature survey many recent studies indicated that the tagging systems are vulnerable to tag spam and malicious users generate the incorrect or misleading tags to confuse the normal participants in the system. For instance, some attackers may repeatedly annotate some images in Flickr with the incorrect tags; so that the normal users, without sufficient knowledge about other participants, may be mislead to open an undesirable image.

1.1 Trust Modeling

Proposed social tagging system makes use of trust modeling algorithms for classification of spam and legitimate texts. Authors in paper [6] proposed that

spam or noise can be injected at three different levels: spam content, spam tag-content association, and spammer. Trust modeling can be performed at each level separately or different levels can be considered jointly to produce trust models. For example, to assess a user's reliability, one can consider not only the user profile, but also the content that the user uploaded to a social system. We categorize trust modeling approaches into two classes according to the target of trust, i.e. user trust modeling and content trust modeling. Content trust modeling is used to classify content like posts, images, and web pages as spam or legitimate. In this category, the target of trust is content, and thus a trust value is given to each content based on its content and/or associated tags. User trust modeling is of two types: static and dynamic.

1.2 Problem Statement

Tagging systems are known to be vulnerable to tag spam. These systems depend on user-generated content, making them both extremely dynamic and tempting targets for spam. So the increasing interest in tagging systems also increases danger from spam. Sometimes the shared content is assigned with inappropriate tags for several reasons. Users are human beings and may commit mistakes but it happens rarely. Most of the time, spammers provide wrong tags on purpose for their advertisement, self-promotion, or to increase the rank of a particular tag in search engines. If this kind of spam is left unchecked, could harm the system in many ways, such as resource sharing openness, information retrieval effectiveness and user experience, etc. One of the major issues in tagging is to identify the most appropriate tags for given content, and at the same time, to eliminate noisy or spam tags. Thus, spam-fighting mechanisms need to be developed to combat the flexible strategies of spammers.

Here we try to understand the problem better, to examine to what extent tagging systems can be manipulated by spammers and to try to devise schemes that may fight spam.

1.3 Review Of Literature

We studied the concepts of using machine learning techniques and classifiers which are used for email spam filtering. As there are various issues related to social tagging systems, we applied all these techniques to build the proposed system which specifically try to combat spam in social tagging systems and try for very efficient results.

This article [1] surveys three categories of potential countermeasures those based on detection, demotion, and prevention. Detection-based strategies attempt to identify spam and remove it or reduce its prominence. Demotion-based strategies attempt to lower the ranking of spam in ordered lists. Prevention based strategies attempt to make contribution of spam more difficult by changing interfaces or limiting user action.

In this paper [2] proposed system uses TrustRank for Combating Web spam. Search engines are today combating web spam with a variety of ad hoc, often proprietary techniques. This paper introducing a comprehensive solution to assist in the detection of web spam. Experimental results show that we can effectively identify a significant number of strongly reputable (non-spam) pages.This paper [3] defines an ideal tagging system that combines legitimate and malicious tags. This model allows studying a range of user tagging behaviors, including the level of moderation and the extent of spam tags, and comparing different query answering and spam protection schemes.

In paper [4] authors proposes the SocialTrust framework for tamper resilient trust establishment in online communities. Social-Trust provides community users with dynamic trust values. Authors experimentally evaluate the SocialTrust framework using real online social networking data consisting of millions of MySpace profiles and relationships.

In this paper [5], authors introduce a set of initial features that can be used for spam classification. These features are evaluated with well-known classifiers (SVM, Naive Bayes, J48 and logistic regression) against a simple baseline of representing a user by the usage of tags.

This paper [6] surveys recent advances in techniques for combating such noise and spam in social tagging. Author proposed a model of a social tagging system that consists of users, contents and tags. Also classified existing studies in the literature into two categories, i.e., content and user trust modeling. Representative techniques in each category were analyzed and compared.

In this paper [7], authors make a contribution towards the development of a privacy-preserving collaborative tagging service, by showing how a specific privacy-enhancing technology, namely tag suppression, can be used to protect end-user privacy. For removing spam author uses Naive Bayes classifiers which is most successful known algorithms for learning to classify text documents.

In this paper [8], the system that classified spam mail and other mail (regular mail) was constructed by two filters with Bayesian theory and SVM(Support Vector Machine) used well by the text classification task as a text classification algorithm. In this paper [9] authors construct different filters using three types of classification, including Naïve Bayes, SVM, and KNN.Naïve Bayes spam email filters are a well-known and powerful type of filters. In this paper [10], authors surveyed social tagging with respect to different aspects. They discussed different user motivations and different ways of tagging web objects. They presented a summary of the various tag generation models. In paper [11] Authors examines the effectiveness of statistically-based approaches

Naive Bayesian anti-spam filters, as it is content-based and self-learning (adaptive) in nature.

II. PROPOSED SYSTEM

2.1. System Framework

The proposed system has been designed with an objective to minimize spam tagging and posting in social networking scenario. The two views for the system would be: admin and user.

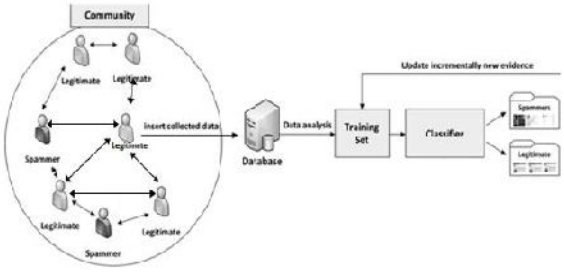

Figure 1: System framework

An admin is a person with complete access to the system. The admin panel would be facilitated with vital features to regulate the postings made by the user. Likewise, a reporting feature would assist him to analyze the spam users in the system.

Admin view would have the right to define the stop words and base words in the application. A stop word can be defined as a word which doesn't affect the logic of the statement and is present in the sentence for grammatical fulfillment eg. a, the, is, that etc. On the other hand, a base word can be defined as a word which directly impacts the logic of the statement e.g. reach, jump, travel, drive etc. Stemming describes the process of transforming a word into its root form.

A user will have the feature to register itself independently on the website by providing essential details suggested by the application. A user can log into the system by facilitating essential login credentials in the application. A user would have a feature to add other users as friends in the system.

A user can add new posts in the application which gets linked to his profile. The system checks for spam (Content Analysis) based on the tags selected by the user for making the post. Depending upon the relation and relevancy of the used tags and the content, the system approves it. Likewise, the system also checks the user trust - static based on the history of posts that he had made in the system. It helps the system to comprehend the nature of the user and its subsequent classification. Also, it considers the posts made by the friends of the users. These factors help to assign a weight of trust to the users' profile and his posts.

Based on all the three checks above, the system fathoms whether the post is a spam or not. The decision is automated and system uses the base words and stop words defined by the admin in his login. A user is permitted to make only limited posts. Once he crosses the threshold, the system blocks the user. A user also has the feature to search for posts made by other users and vote up or vote down to them.

An admin sees a list if users which are blocked over a period of time and also checks the results generated by the system Bayesian spam filtering algorithm. It makes use of a naive Bayes classifier on bag of words features to identify spam tag, an approach commonly used in text classification. Naive Bayes classifiers work by correlating the use of tokens (typically words, or sometimes other things), with spam and non-spam tags and then using Bayesian inference to calculate a probability that a tag is or is not spam.

Our system that classified spam tags and regular was constructed by two filters with Bayesian theory and KNN used well by the text classification task as a trust modeling algorithm. Trust modeling algorithms takes known set of input data and known responses to the data as output, and trains a model to generate reasonable predictions for the response to new data.System framework is described to demonstrate how a social tagging system can benefit from trust modeling with the adaptation of classification algorithms.

The proposed system framework requires training of keywords that can be provided by a previous set of spam and legitimate messages. It keeps track of each word that occurs only in spam, only in legitimate messages, and in both. Based on these word occurrence statistics also called tokens, incoming unseen messages are processed and classified accordingly.

2.2. System Model

System model is described to demonstrate how a social tagging system can benefit from trust modeling with the adaptation of classification algorithms. System model consists of following steps:

- Steps:
  1. Training data of annotated tags
     - In training data set I want to introduce some tag examples to demonstrate process of Navie Bayes classification.
  2. A set of classes
     - In our case two possible classes
     - Can further be personalized
  3. Feature Extraction
     - Tokenization
     - Domain specific features
     - Most often features to be selected
  4. Classify (each message)
     - Calculate posterior probabilities
  5. Evaluate results

2.2.1. Spam Filtering Method:

i. Naive Bayes classifiers

Naive Bayes classifier is one of the most successful known algorithms for learning to classify text

documents. Bayesian spam filtering has become a popular approach to distinguish spam texts from legitimate texts. The filter doesn't know probability of new word in advance, and must first be trained so it can build them up. Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. To train the filter, the user must manually indicate whether a new tag is spam or not. For all words in each training posts, the filter will adjust the probabilities that each word will appear in spam or legitimate keywords in its database. The naive Bayes classifier's beauty is in its simplicity, computational efficiency, and good classification performance.

Let Pr(S) be the probability that a message is spam which is the total number of spam messages divided by the total amount of messages. Now the goal would to be taking a feature of word that describes a spam message and calculate the probability of that.

$$P_r(S|W) = \frac{Pr(W|S) \cdot Pr(S)}{Pr(W|S) \cdot Pr(S) + Pr(W|H) \cdot Pr(S)}$$

Here Pr(S|W) is the probability that a message is a spam, knowing that the word is in it;Pr(S) is the overall probability that any given message is spam; Pr(W|S) is the probability that the word appears in spam messages;Pr(H) is the overall probability that any given message is not spam (is "ham"); Pr(W|S) is the probability that the word appears in ham messages [21].

For evaluating result consider the message m and determine the value for feature w. Then take the calculated probabilities and calculated the probability that it is spam and the probability that it is legit. Compare those two probabilities to classify the message as spam or legit.

ii. KNN classifier

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). One advantage of this algorithm is that there isn't really a training phase. However, for classification of a message, all distances between that message and all the training examples must be calculated and the knearest neighbors need to be found and counted.

- **Algorithm**

Assumption: there are previously minimum 10 post detected.

1. define k=5,
2. Identify parameter for input post.
3. Compare with all spam post.
4. Calculate distance of current post with other spam post using formula:-
5. x=avg(total count for all keywords for that post)
6. y=avg(support value)
7. d= sqrt((x1-x2)^2 + (y1-y2)^2 + (z1-z2)^2 ….)
8. Store d in array, with post id.
9. end for loop.
10. sort distance array
11. Identify top k value.
12. Calculate avg between these 5 distance value.
13. this is k distance for spam.
14. repeat above step for valid post.
15. identify k distance which is less.
16. return less=spam

For evaluating result consider a message m, find the k nearest neighbors and count the number of each label whether spam or not that are given from the neighbors. If there are more spam messages in the k nearest neighbors then it is classified as spam. If not, then that message is classified as legitimate tag.

III. RESULT ANALYSIS

System shows classification of spam (bad) posts from total posts for both the algorithms on dashboard as shown in figure 3.Then figure 4 shows classification of tags separately for each users using KNN classifier.

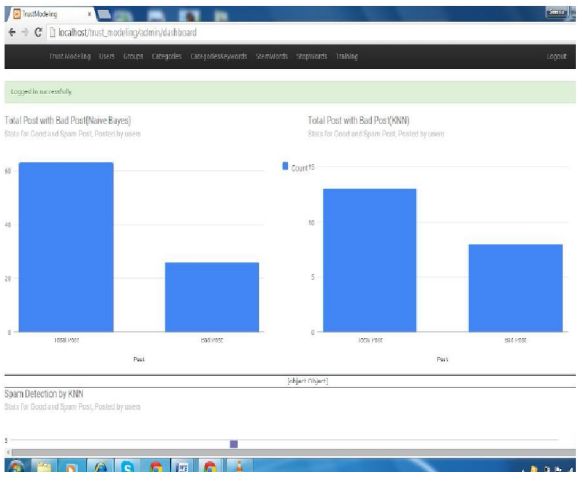

Figure 2: Classified posts for both algorithms

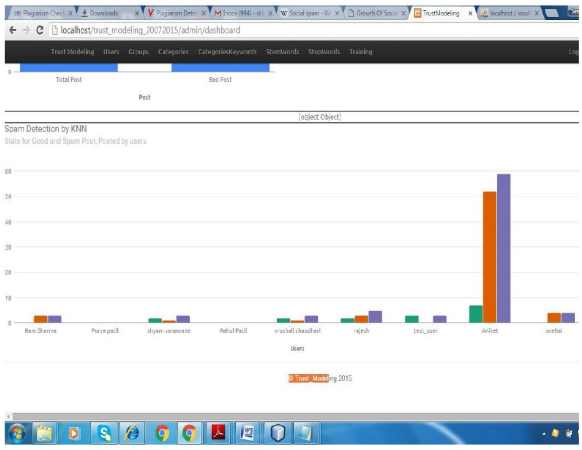

Figure 3: Classified posts using KNN algorithm

CONCLUSIONS

Here we dealt with one of the key issues in social tagging systems: combating noise and spam. Based on categorization of trust modeling i.e content and user, we focused on classification of contents of social tagging system as spam or legitimate. In a social tagging system, spam or noise can be injected at three different levels: spam content, spam tag-content association, and spammer. Trust modeling can be performed at each level separately or different levels can be considered jointly to produce trust models. Trust Modeling is one of the current techniques for noise and spam reduction focus only on textual tag processing and user profile analysis. Proposed model is described to demonstrate how a social tagging system can benefit from trust modeling with the adaptation of classification algorithms.

The proposed system has been designed with an objective to minimize spam tagging and posting in social networking scenario. The system checks for spam (Content Analysis) based on the tags selected by the user for making the post, history of posts and user profile. Here we use Naive Bayesian Model for developing proposed system to filter spam tags in social tagging systems. Then we use KNN classifier for spam filtering and check the result. Comparative study states that KNN gives better result than NB classifier.

ACKNOWLEDGMENTS

We would like to thank Prof.Dipti Patil for helpful discussions. We would also like to thank the anonymous reviewers for their helpful suggestions.

REFERENCES

[1] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Fighting spam on social Web sites: A survey of approaches and future challenges," IEEE Internet Comput., vol.11, no. 6, pp. 36–45, Nov. 2007.

[2] Z. Gyongyi, H. Garc ia-Molina, and J. Pedersen, "Combating Web spam with TrustRank," in Proc. VLDB, Aug. 2004, pp. 576–587.

[3] G. Koutrika, F. A. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina, "Combating spam in tagging systems: An evaluation," ACM TWEB, vol. 2, no. 4, pp.22:1–22:34, Oct. 2008.

[4] J. Caverlee, L. Liu, and S. Webb, "SocialTrust: Tamperresilient trust establishment in online communities," in Proc. ACM JCDL, June 2008, pp.104–114.

[5] B. Krause, C . Schmitz, A. Hotho, and G. Stum, "The antisocial tagger: Detecting spam in social bookmarking systems," in Proc. ACM AIRWeb, Apr. 2008,pp. 61–68.

[6] I.Ivanov, P. Vajda, J. S. Lee, and T. Ebrahimi, "In tags we trust: Trust modeling in social tagging of multimedia content," IEEE Signal Proc. Mag., vol. 29, no. 2, pp. 98–107, Mar. 2012.

[7] L.Sundarrajan, S.Gunasekaran "Social Networks Privacy-Preserving On Collaborative Tagging and Spam Filter Using Naive Bayes Algorithm" IJIRCCE, Vol. 2, Issue 10, October 2014.

[8] Ayahiko Niimi, Hirofumi Inomata, Masaki Miyamoto and Osamu Konishi "Evaluation of Bayesian Spam Filter and SVM Spam Filter"2004.

[9] Yun-Nung Chen, Che-An Lu, Chao-Yu Huang "Anti-Spam Filter Based on Naïve Bayes,SVM, and KNN model",AI TERMPROJECT,2009.

[10] Manish Gupta, Rui Li, Zhijun Yin, Jiawei Han "An overview of social tagging and Applications",Springer International Publishing, Mar. 2011.

[11] Vikas P. Deshpande, Robert F. Erbacher, and Chris Harris "An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques", IEEE, June 2007.

[12] Flickr Web site.[Online]. Available:http://www.flickr.com

[13] Facebook Web site. [Online] .Available: http://www.facebook.com

[14] Delicious Web site. [Online]. Available: http://www.delicious.com

[15] eBay Web site. [Online]. Available: http://www.ebay.com

[16] Amazon Web site. [Online]. Available: http://www.amazon.com

[17] Epinions Web site. [Online]. Available: http://www.epinions.com

[18] Twitter Web site. [Online]. Available: http://www.twitter.com

[19] Panoramio Web site. [Online]. Available: http://www.panoramio.com

[20] MySpace Web site. [Online]. Available: http://www.myspace.com

[21] http://en.wikipedia.org/wiki/Bayesian_spam_filtering

[22] http://techcrunchies.com/growth-of-social-media-spamstatistics-for-2013

[23] Paul Graham: A Plan for Spam, http://www.paulgraham.com/spam.html

[24] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[25] YouTube Web site. [Online]. Available: http://www.youtube.com

★★★