2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery

Detecting Tag Spam in Social Tagging Systems with Collaborative Knowledge

Kaipeng Liu
Research Center of Computer Network and Information Security Technology
Harbin Institute of Technology
Harbin, China
liukaipeng@pact518.hit.edu.cn

Binxing Fang
Information Security Research Center
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
bxfang@ict.ac.cn

Yu Zhang
Research Center of Computer Network and Information Security Technology
Harbin Institute of Technology
Harbin, China
zhangy@pact518.hit.edu.cn

***Abstract*—Social tagging systems allow collaborative users to annotate shared resources with tags. Since they rely on user-contributed content, social tagging systems are vulnerable to spam annotations, which are generated by malicious users to mislead or confuse legitimate users. Thus, mechanisms for spam detection need to be developed to combat the flexible strategies of spammers for the success of social tagging systems. Since annotations are lack of relevant feature, the classical method of training classifier to detect spam is hard to implement. However, with their collaborative nature, knowledge on the tagging scheme do exists in the way numerous participants annotating resources with tags. In this paper, we propose a simple but remarkably effective approach for detecting tag spam in social tagging systems with collaborative knowledge. We harness the wisdom of crowds to discover the knowledge on what should be high quality annotations for resources. This knowledge is then used to tell spam posts from the legitimate ones. A distinct feature of our approach is that, it can be easily extended for user level spam detection and can do well in both levels. The proposed approach is evaluated on data set collected from real-world system. Experimental results show a convincing performance of proposed approach.**

***Keywords*-social tagging; tag spam; collaborative knowledge;**

I. Introduction

Social tagging systems allow users to submit shared resources and to annotate them with descriptive tags collaboratively, forming the so-called *folksonomies*. These systems provide powerful infrastructure for semantic annotation and information sharing, promoting various kinds of Internet-based activities such as web exploration and community formation. Take the social bookmarking systems as an example, instead of keeping a local copy of favorite URLs, users can store and access their bookmarks online through a web interface. The underlying infrastructure then makes all stored information sharable among users, allowing for improved information retrieval and forming communities among users with similar interests. The success of these systems is mainly relying on the easy-to-use system interface and immediate benefits yielding from system without too much overhead.

However, as any other systems that rely on user-generated content, social tagging systems are vulnerable to spam. As we know, the Web is vulnerable to search engine spam, which is content created to mislead search engines into promoting the ranking of some websites to a certain level that is higher than they deserve [1]. Web spam becomes a big problem for search engines nowadays. Spammers choose search engine as their attack targets because many people use search engines as entryways to the Web. In an analogous fashion, with the rapidly increasing of popularity, social tagging systems are becoming new entryways to various shared resources such as bookmarks, photos and videos, etc. This success also attracts spammers' attention to consider social tagging systems as a new platform to publish their content: all they need is an account; then they can freely post entries pointing to the target websites. Often the spam is generated for advertising, self-promotion, or promoting the visibility of the target resources. This kind of spam, if left unchecked, could harm the system in many ways, such as resource sharing openness, information retrieval effectiveness and user experience, etc. Thus, spam-fighting mechanisms need to be developed to combat the flexible strategies of spammers.

There are different kinds of spam existing in social tagging systems such as resource spam and tag spam. Resource spam refers to the resources posted to the system which the legitimate users do not wish to share, while tag spam is the content posted and tagged in a way to mislead or simply confuse other users. Notice that these kinds of spam are not exclusive, they may coexist together. For example, a spammer who posts a gambling website for advertising, may also use attractive but misleading tags like "cash" and "earn money". We will focus on detecting tag spam in this study, while the methodology is not necessarily restrict to the scope of tag spam detection.

Several challenges exist while developing spam detection mechanisms in social tagging systems. Since annotations consist of simply resources and descriptive tags, relevant features are hard to be selected for a classification algorithm to work. It is also difficult to build the training data sets manually. Thus, the classical method of training classifier to detect spam is difficult to implement in social tagging systems.

In this paper, we propose a simple but remarkable effective approach to detect tag spam in social tagging systems

978-0-7695-3735-1/09 $25.00 © 2009 IEEE
DOI 10.1109/FSKD.2009.401

by harnessing the wisdom of crowds. We uncover the underlying tagging scheme in folksonomies to discover the knowledge on what are high quality annotations for resources. Then, we use this knowledge to tell the low quality spam annotations from the legitimate ones. Moreover, we extend this method to user level for detecting spammers that post spam content frequently.

The remainder of the paper is structured as follows. Section II discuss the related work on combating spam in social tagging systems. The details of the proposed approach are presented in Section III. Section IV gives the experimental evaluation of the proposed approach. We conclude this paper and discuss some issues with our approach in Section V.

II. Related work

A. The impact of social tagging spam

Several research has been conducted on studying the impact of spam on social tagging systems. Georgia Koutrika et al. [2], [3] study the impact of spamming through a framework for modeling social tagging system and user tagging behavior. They also describe a method for ranking documents matching a tag based on taggers reliability. They develop a system model consists of resources, tags and users. A particular instance of the tagging system is populated with different user tagging behavior models, including the good user model, bad user model, targeted attack model and several other models. They evaluate the impact of tag spam with a metric called SpamFactor with different search models, including boolean search, occurrence-based search and coincidence-based search. Several conclusions are derived based on the experiments with synthetic models.

Paul Heymann et al. [4] survey the approaches to combating spam in social websites. They discuss the characteristics of social websites, including one controlling entity, well-defined interactions, identity and multiple interfaces, which substantially change the relationship between service providers and spammers. Based on this discussion, they survey three categories of potential countermeasures, including those based on detection, which involves spam user identification; demotion, which refers to designing algorithm to reduce the prominence of spam content; and prevention, which are methods to make contributing spam content difficult. The method for evaluating these countermeasures is also studied in their work.

B. Detecting social tagging spam

Besides the above modeling analysis and methodology discussion on combating spam in social tagging systems, several approaches for detecting spam in real world systems are also proposed in the literature. Beate Krause et al. [5] transfer the machine learning approach to a social bookmarking setting to identify spammers. They present features considering the topological, semantic and profile-based information which people make public when using the system. The experiments on the data set of a snapshot of the social bookmarking system BibSonomy give a comparison of the performance of a large set of different classification models. Their work presents a groundwork for building of more elaborate spam detection mechanisms.

Tonie Bogers et al. [6] use language models to detect spam in social bookmarking systems. Their method is based on the intuitive notion that similar users and posts tend to use the same language. To detect spam users in the system, they use the users and posts that are most similar to the incoming users and their posts to determine the spam status of these new ones. At first, they rank all users in the system by KL-divergence of the language models of their posts and the language model of the new post or user. Then, they predict a spam label for the new user by looking at the spam labels assigned to the most similar users in the system. They also give a comparison of using language models at two different levels of granularity: individual posts and individual users. Experiments with the data set of BibSonomy show that their method could achieve a high AUC score. Anestis Gkanogiannis et al. develop a supervised learning algorithm and apply it to the spam detection problem [7]. Naive Bayes classifier is also used for spam detection by Chanju Kim et al. [8].

III. Method

As mentioned in Section I, the classical method for detecting spam with trained classifier is hard to implement in social tagging systems. However, if we take a deeper step towards the collaborative nature of social tagging systems, we will find that knowledge on the tagging scheme do exists in the way numerous users annotating resources with tags. As Halpin et al. have pointed out [9], if there are sufficient active users, over time a stable distribution with a limited number of stable tags and a much larger "long-tail" of more idiosyncratic tags will develop. This stabilized distribution (see Section IV-A) can be considered as a convergence of tagging scheme. Thus, with collaborative knowledge that is implicitly present in the folksonomies, we may develop a effective way to detect tag spam in social tagging systems.

A. Post level spam detection

We formally define a folksonomy as a tuple with four components $\mathcal{F} = (U, T, R, A)$, where U, T, and R are finite sets of users, tags and resources, respectively, and A is a ternary relation between them, i.e., $A \subseteq U \times T \times R$, whose elements are called tag assignments. A tag assignment $a = (u, t, r) \in A$ represents user u annotates resource r with tag t. We also denote all the posts in the folksonomy with $P \subseteq U \times 2^T \times R$. A post $p = (u, T_p = \{t|(u, t, r) \in A\}, r)$ denotes an actual post in the system for user u annotating resource r with a set of tags T_p.

With the convergence of tag usage, the knowledge on tagging scheme to resources develops. We represent this

knowledge on a resource r as a tagging scheme vector S_r, for which the index $S_r(t)$ is the number of times the tag t is used by separate users to annotate the resource r,

$$S_r(t) = |U_r^t|, \tag{1}$$

where U_r^t is the set of users who have posted to the resources r with tag t. To measure how much information can be gained by a tag assignment, the tag t's tagging information value $V_r(t)$ with respect to resource r is defined as follows,

$$V_r(t) = \frac{S_r(t)}{\sum_{t' \in T} S_r(t')}. \tag{2}$$

Then, the posting information value $V(p)$ of post $p = (u, T_p, r)$ can be calculated as the average of the tags' tagging information values in T_p,

$$V(p = (u, T_p, r)) = \frac{\sum_{t \in T_p} V_r(t)}{|T_p|}. \tag{3}$$

As we can see from (3), the posting information value represents the knowledge embedded in the post p. A post of low information value indicates a divergence from crowds and a poor value of tagging information. Thus, it can be used to tell the low quality spam posts from legitimate ones. However, since this value may vary when the corpus changes, its absolute value has less meaning. Thus, while identifying spam posts, instead of using a threshold value to classify the post once and for all, we employ a iterative manner to pick out spam posts gradually. The complete algorithm is shown in Algorithm 1.

Algorithm 1 Algorithm for detecting spam posts

```
1:  procedure DETECTSPAMPOSTS(V_min, F_max)
2:    P_spam ← ∅
3:    while |P_spam|/|P| ≤ F_max do
4:      for all r ∈ R do
5:        Get tagging scheme vector S_r with (1)
6:        for all t ∈ T do
7:          Get tagging information value V_r(t) with (2)
8:        end for
9:      end for
10:     for all p ∈ P do
11:       Get posting information value V(p) with (3)
12:       if V(p) < V_min then
13:         P_spam ← P_spam ∪ {p}
14:         A ← A \ {(u, t, r)|t ∈ T_p}
15:       end if
16:     end for
17:   end while
18:   return P_spam
19: end procedure
```

We first calculate the posting information values for each post, and identify the spam posts according to V_{min}, which is a threshold value for the minimum information value of legitimate posts (lines 4–16). Then, we eliminate all the tag assignments related to these spam posts in the set of annotations A (line 14). The post information values are recalculated, and the spam posts are identified again. This progress is repeated until there is no post with a posting information value lower than V_{min}, or the fraction of detected spam posts exceeds F_{max}, which is a predefined value for the maximum fraction of posts to be identified as spam.

B. User level spam detection

We can extend the above method for detecting spam posts to the user level. A straight forward idea is a to use the average posting information value to represent the quality of users' posting profile. However, this method does not take the resources' importance into account. Since the popularity of resources in social tagging systems directly reflect the collaborative knowledge on resources' quality, we can use this knowledge as the measure of their importance. Specifically, we calculate the resource r's importance $I(r)$ as follows,

$$I(r) = \frac{|U_r|}{\sum_{r' \in R} |U_{r'}|}, \tag{4}$$

where U_r is the set of users who have posted to resource r.

Then, a user u's posting quality $V(u)$ can be calculated as the weighted average of posting information value, with resource importance as weight,

$$V(u) = \frac{\sum_{p=(u,T_p,r) \in P_u} I(r)V(p)}{|P_u|} \tag{5}$$

where P_u is the set of user u's posts. We also use a user u's posting information loss $L(u)$ to measure the harm he/she has done to the system,

$$L(u) = \sum_{p=(u,T_p,r) \in P_u} I(r)(1 - V(p)). \tag{6}$$

We use both posting quality and information loss to identify spammers with an analogous method to the one for detecting spam posts described above.

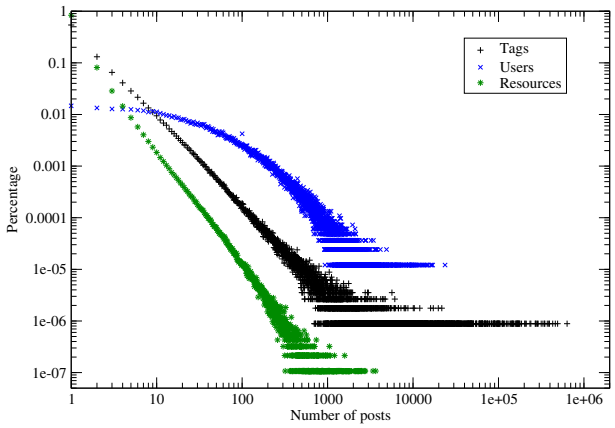
IV. Experiments

In this section, we evaluate our approach on the data set collected from real-world social tagging system. We describe the data set below in Section IV-A, and present the experimental results in Section IV-B.

A. Data set

The data set used in this paper is a partial dump of Delicious[1] representing post activities during a limited period of time. We crawl post pages from Delicious and extract post information to build the data set (see Table I). Figure 1

[1] http://delicious.com



]
]

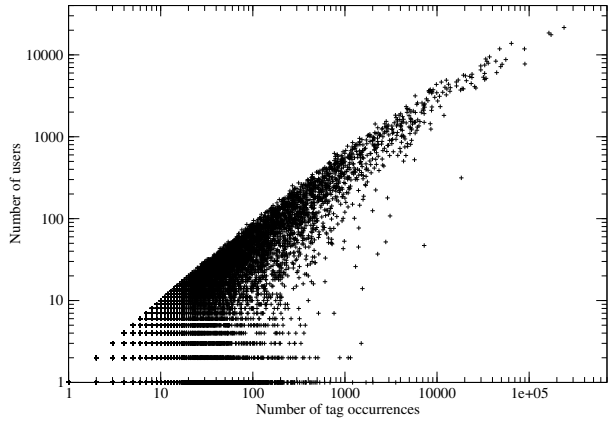


Figure 2. The correlation between the number of tag occurrences and its popularity in Delicious.

shows the distributions of users, tags and resources for posts in the data set. The plotted points of tag distribution and resource distribution in the figure are nearly in a straight line on the log-log scale, indicating that these distributions follow the power law. Our crawling strategy tends to ignore those users who post seldomly, leading to the divergence of user distribution in this figure. The scale-free power law distribution for tagging frequency suggests that once one version of a tag becomes very popular, it is used even more in the future. Unless a new tag with more information value is discovered, further tagging will only reinforce the preexisted tagging scheme [9]. Figure 2 shows the strong correlation between the number of users using a particular tag and its popularity. Both figures suggest that there is a strong sense of agreement on common tag usages among users.

B. Results

For post level spam detection, we use Algorithm 1 to generate a list of identified spam posts, and manually examine the first 2,000 posts in the list. To gain a insight into the structure of spam posts, we also classify the detected spam posts into several categories by their characteristics. The results are shown in Table II.

Table I
THE DATA SET OF DELICIOUS.

#Users	#Tags	#Resources	#Posts
82,541	1,129,656	9,321,338	17,387,756

Table II
CATEGORIZATION OF IDENTIFIED POSTS.

Category	Number of posts	Percentage
Advertising	1072	53.60%
Misleading	545	27.25%
Misused	258	12.90%
Non-English	35	1.75%
Legitimate	78	3.90%
Inaccessible	12	0.60%

From Table II we can see that, most of the spam posts are submitted for advertising. Spammers use tags like “free”, “cash” and “offers”, etc., to annotate their target websites, while other users pay more attention to the actual content. For example, a website for football betting is tagged by spammers with tag “earn money”, while most other users prefer “football betting”. Misleading tags are those attractive tags used by the spammers to gain attention from other users, while the tags themselves have nothing to do with the resources’ content. For example, a website for file sharing is tagged with “sex” and “adult” by spammers. In fact, the advertising and misleading posts are not exclusive. The results present here are just roughly categorized. Misused tags are generated by incautious or uncommon users. Such tags include misspelled words, artificial words for personal use and odd words hard to understand. Since the majority of users in Delicious use English while annotating resources, some non-English tags are also (mis)identified by our algorithm. One could argue that these posts should not to be considered as spam. We do not study multilingual problem in this work, so we leave this issue open. Unfortunately, some legitimate posts are misidentified. The reason is that, for those unpopular resources, the collaborative knowledge is not developed yet. Even a few incorrect annotations may dominate the lesser correct ones. Thus, our algorithm will catch the legitimate ones as spam. With the limitation of collaborative knowledge, we can not settle this problem easily.

For user level spam detection, we sort the detected spammers by posting information loss in descending order and manually examine the top 500 users. The categorization results are shown in Table III. These results are similar to those for post level spam detection. As we get a high

Table III
CATEGORIZATION OF IDENTIFIED USERS.

Category	Number of users	Percentage
Ad. & Misleading	457	91.4%
Misuse	24	4.8%
Non-English	11	2.2%
Legitimate	8	1.6%

precision of 93.75% in the top 2000 posts and 96.20% in the top 500 users, we can say that our approach can do well in both post and user level spam detection.

V. CONCLUSION AND DISCUSSION

This paper has demonstrated our approach to detect tag spam in social tagging systems. We uncover the underlying tag scheme that is already presented in the way collaborative users annotating resources with tags. This collaborative knowledge is then used to measure the quality of posts and tell spam posts from legitimate ones. A iterative spam detection algorithm is developed to identify spam posts by their information value. This method can be also extended to user level for detecting spammers. We have done experiments on data set collected from real-world system, and the experimental results show a convincing performance of our approach.

One issue to discuss is that while the collaborative knowledge in the folksonomies is considered for detecting spam, the knowledge in *personomies* is ignored. The personomy $\mathcal{P}_u$ of a given user $u \in U$ is the restriction of $\mathcal{F}$ to u, i.e., $\mathcal{P}_u = (T_u, R_u, P_u)$, which represents the user's personal preference of posting and tagging scheme. If the user's behavior heavily diverge from the others in the system, while legitimate, he/she might be identified as a spammer by our approach. We could argue that this is the right way, since from the other users' point of view, his/her information is lack of value. On the other hand, if we can take the personomies into account while identifying spammers, it will be a great improvement to our approach. We will work on this in the future.

REFERENCES

[1] Z. Gyöngyi and H. Garcia-Molina, "Web spam taxonomy," in *AIRWeb 2005, First International Workshop on Adversarial Information Retrieval on the Web, Chiba, Japan*, 2005, pp. 39–47.

[2] G. Koutrika, F. A. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina, "Combating spam in tagging systems," in *AIRWeb 2007, Third International Workshop on Adversarial Information Retrieval on the Web, Banff, Canada*, 2007, pp. 57–64.

[3] ——, "Combating spam in tagging systems: An evaluation," *TWEB*, vol. 2, no. 4, 2008.

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

[5] B. Krause, C. Schmitz, A. Hotho, and G. Stumme, "The anti-social tagger: detecting spam in social bookmarking systems," in *AIRWeb 2008, Fourth International Workshop on Adversarial Information Retrieval on the Web, Beijing, China*, 2008, pp. 61–68.

[6] T. Bogers and A. van den Bosch, "Using Language Models for Spam Detection in Social Bookmarking," in *Machine Learning and Knowledge Discovery in Databases, ECML/PKDD 2008, Proceedings, Part II*, 2008.

[7] A. Gkanogiannis and T. Kalamboukis, "A novel supervised learning algorithm and its use for Spam Detection in Social Bookmarking Systems," in *Machine Learning and Knowledge Discovery in Databases, ECML/PKDD 2008, Proceedings, Part II*, 2008.

[8] C. Kim and K.-B. Hwang, "Naive Bayes Classifier Learning with Feature Selection for Spam Detection in Social Bookmarking," in *Machine Learning and Knowledge Discovery in Databases, ECML/PKDD 2008, Proceedings, Part II*, 2008.

[9] H. Halpin, V. Robu, and H. Shepherd, "The complex dynamics of collaborative tagging," in *Proceedings of the 16th international conference on World Wide Web*. ACM New York, NY, USA, 2007, pp. 211–220.