Spam Discovery In Social Bookmarking Sites

Chintan Sheth
Department of Computer Science
School of Informatics
Indiana University
Bloomington, IN 47408, USA
cjsheth@indiana.edu

Deepak Konidena
Department of Computer Science
School of Informatics
Indiana University
Bloomington, IN 47408, USA
bkoniden@indiana.edu

Fil Menczer
Department of Computer Science
School of Informatics
Indiana University
Bloomington, IN 47408, USA
fil@indiana.edu

ABSTRACT
With a growing popularity of social bookmarking sites and also due to the community structure of bookmarking sites, spammers found new techniques to attack the web. To retain the benefits of sharing one's web content, spam – discovery mechanisms that can face the flexible strategies of spammers need to be developed. In this paper we use a supervised learning approach to build a classifier that detects user spam. We explore various relationships between tags, bookmarks and users to detect spammers' activity. We will present features considering this semantic-behavior and user activity. The dataset used is a snapshot of Social Bookmarking system BibSonomy. For classification we are using an open source SVM library libsvm. Our results present a starting point of how we can detect spam in social bookmarking sites.

Categories and Subject Descriptors
H.3 [Information Storage and Retrieval]: Miscellaneous

General Terms
Algorithms, Experimentation

Keywords
spam detection,

1.INTRODUCTION
Social bookmarking is a method for Internet users to store, organize, search, and manage bookmarks of web pages on the Internet with the help of metadata, typically in the form of tags that collectively and/or collaboratively become a folksonomy. Folksonomy is also called *social tagging*, "the process by which many users add metadata in the form of keywords to shared content".[9]

Social Bookmarking systems have become an attractive place for posting web spam. These systems allow users to annotate and share bookmarks, Spammers misuse the popularity and the high PageRank of social bookmarking systems for their purposes.

In order to retain the original benefits of social bookmarking systems, techniques need to be developed which prevent spammers from publishing content in these websites[1]. The problem can be considered as a binary classification task. Based on different features that describe a user and his semantic-behavior, a model is built to classify unknown samples either as spam or non-spam .

This classification problem consists of two steps. The first one is to select features for describing the user activity. The second step is selection of the appropriate classifier for the problem. In this paper we would be using SVM classifier.

This article is organized as follows. In Section 2 we present the Dataset Analysis. In Section 3 we provide the feature selection. be discussing

2. DATASET

Bibsonomy is a social bookmarking site that allows users to annotate and share bookmarks and publication references. Its labeled dataset is used for experimentation in this paper. It consists of users, tags, bookmarks, bibtex and a user's profile information until the end of 2008. The Figures below indicate a snapshot of the original dataset along with the downsized dataset that we used for training and testing purposes.

Table 1: Original dataset

	Users	Tags	Bookmarks
Spam	29248	13258759	2059991
Non Spam	2467	816197	181833

Table 2 : Training and Test dataset

	Users	Tags	Bookmarks
Train	1299	474465	72172
Test	433	131186	24088

3. FEATURE SELECTION
The first step in any classification problem is to identify features that would help the classifier in deciding if a particular user is a spammer or not. After performing an exhaustive analysis of the data , we found 7 distinct features that could differentiate the activity of a spammer to a normal user. An instance in the training or test data , that is fed into the classifier , can be seen as a vector with a class label followed by a feature and its corresponding value.

An extensive study of the data revealed the following features:

2.1. Count of Special Characters in Tags :- Special characters and digits [1-9] more oftenly occurred in spammer's tags.

Stats : Spammers (471/912) and Non-spammers (423/822).

2.2. Count of Bag of Words :- We collected a list of words which more frequently appeared in spammer's tag list than in a non-spammer's one. We used term frequency as the measure , and collected 50 most frequent spam tags.

Stats : Spammers (359/912) and Non-spammers (124/822).

2.3. Tag Length :- We observed that a significant number of spammers preferred shorter tags (one letter or two at the most).

Stats : Spammers (415/912) and Non-spammers (209/822).

2.4. Count of No.of Tags Per Day Per User :- Spammers , in our study, happened to tag higher number of tags / day when compared to non-spammers.

Stats : Spammers (77) and Non-spammers (65)

2.5. **Count of Number of Bookmarks Per User :-** Spammers on the whole tagged to more number of bookmarks as compared to non – spammers.

2.6. Number of Bookmarks/Tag/User :- Average number of bookmarks per tag is higher for spammers.

2.7. Number of Tags Per Bookmark Per User :- Spammers tag more bookmarks per tag than Non-spammers.

2 EXPERIMENT

We first downsized the dataset to include 912 spammers and 822 non spammers. For downsizing the data we ran perl scripts.

After finding the features we wrote respective perl scripts to validate those features to get an idea of whether those features can build the classifier effectively.

Once we found the features to be valid we made the SVM input file which would be used by the classifier to train the dataset and used a small testing set for finding the accuracy of the system.

3 EVALUATION AND RESULTS

The table below describes the Confusion Matrix of the experiment. We found 155 true positives which is spammers in this case , and 73 false positives which are Non-spammers identified as spammers. The precision and recall scores are 78% and 69% respectively. The accuracy or the F-score is about 73.2%.

We used RBF kernel with five fold cross validation using LibSVM and obtained the below results.

Table 3 : Confusion Matrix

	Spammers	Non Spammers
Spammers	155	73
Non Spammers	44	161

Precision: 78%

Recall: 69%

F-score : 73.2%

ROC CURVE

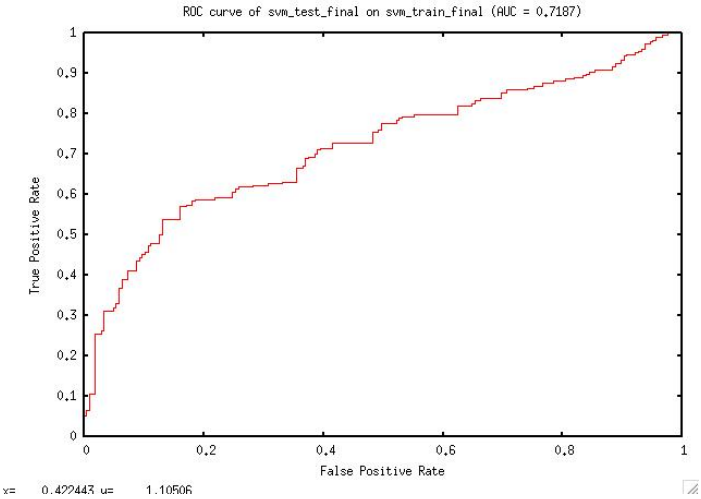

Figure 1 ROC Curve

4 References

[1] B. Krause, Schimitz, A. Hotho, and G. Stumme, "The anti-social tagger - detecting spam in social bookmarking systems," in *Fourth International Workshop on Adversarial Information Retrieval on the Web*, April 2008.

[2]C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors: web spam detection using the web topology," in *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. New York,NY,USA:ACM,2007, pp. 423-430.

[3] Q. Gan and T. Suel, "Improving web spam classifiers using link structure," in *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*. New York, NY, USA: ACM,2007,pp.17-20.

[4]Heymann, G. Koutrika, and H. G. 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] Gkanogiannis, A. & Kalamboukis, T. (2008), A novel supervised learning algorithm and its use for Spam Detection in Social Bookmarking Systems, *in* 'ECML PKDD Discovery Challenge '08' .

[6] Hotho, A.; Jäschke, R.; Schmitz, C. & Stumme, G. (2006), BibSonomy: A Social Bookmark and Publication Sharing System, *in* Aldo de Moor; Simon Polovina & Harry Delugach, ed., 'Proceedings of the First Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures',Aalborg Universitetsforlag,Aalborg,pp.87-102

[7] Toine Bogers and Antal van den Bosch,"Using Language Models for Spam Detection in Social Bookmarking in "ECML PKDD Discovery Challenge '08 '.

[8] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," 2001.

[9] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *J. Inf. Sci.*, vol. 32, no. 2, pp. 198-208, April 2006.