Detecting Spammers in Community Question Answering

Zhuoye Ding, Yeyun Gong, Yaqian Zhou, Qi Zhang, Xuanjing Huang
Fudan University
School of Computer Science
{09110240024,12110240006,zhouyaqian,qz,xjhuang}@fudan.edu.cn

Abstract

As the popularity of Community Question Answering(CQA) increases, spamming activities also picked up in numbers and variety. On CQA sites, spammers often pretend to ask questions, and select answers which were published by their partners or themselves as the best answers. These fake best answers cannot be easily detected by neither existing methods nor common users. In this paper, we address the issue of detecting spammers on CQA sites. We formulate the task as an optimization problem. Social information is incorporated by adding graph regularization constraints to the text-based predictor. To evaluate the proposed approach, we crawled a data set from a CQA portal. Experimental results demonstrate that the proposed method can achieve better performance than some state-of-the-art methods.

1 Introduction

Due to the massive growth of Web 2.0 technologies, user-generated content has become a primary source of various types of content. Community Question Answering (CQA) services have also attracted continuously growing interest. They allow users to submit questions and answer questions asked by other users. A huge number of users contributed enormous questions and answers on popular CQA sites such as Yahoo! Answers[1], Baidu Zhidao[2], Facebook Questions[3], and so on. According to a statistic from Yahoo, Yahoo! Answers receives more than 0.82 million questions and answers per day[4].

On CQA sites, users are primary contributors of content. The volunteer-driven mechanism brings many positive effects, including the rapid growth in size, great user experience, immediate response, and so on. However, the open access and reliance on users have also made these systems becoming targets of spammers. They post advertisements or other irrelevant answers aiming at spreading advertise or achieving other goals. Some spammers directly publish content to answer questions asked by common users. Additionally, another kind of spammers (we refer them as "*best answer spammers*") create multiple user accounts, and use some accounts to ask a question, the others to provide answers which are selected as the best answers by themselves. They deliberately organize themselves in order to deceive readers. This kind of spammers are even more hazardous, since they are neither easily ignored nor identifiable by a human reader. Google Confucius CQA system also reported that best answer spammers may generate amounts of fake best answers, which could have a non-trivial impact on the quality of machine learning model (Si et al., 2010).

With the increasing requirements, spammer detection has received considerable attentions, including e-mails(L.Gomes et al., 2007; C.Wu et al., 2005), web spammer (Cheng et al., 2011), review spammer (Lim et al., 2010; N.Jindal and B.Liu, 2008; ott et al., 2011), social media spammer (Zhu et al., 2012; Bosma et al., 2012; Wang, 2010). However, little work has been done about spammers on CQA sites. Filling this need is a challenging task. The existing approaches of spam detection can be roughly into two directions. The first direction usually relied on costly human-labeled training data for building spam classifiers based on textual features (Y.Liu et al., 2008; Y.Xie et al.,

[1] http://answers.yahoo.com
[2] http://zhidao.baidu.com
[3] http://www.facebook.com
[4] http://yanswersblog.com/index.php/archives/2010/05/03/1-billion-answers-served

International Joint Conference on Natural Language Processing, pages 118–126,
Nagoya, Japan, 14-18 October 2013.

2008; Ntoulas et al., 2006; Gyongyi and Molina, 2004). However, since fake best answers are well designed and lack of easily identifiable textual patterns, text-based methods cannot achieve satisfactory performance. Another direction relied solely on hyperlink graph in the web (Z.Gyongyi et al., 2004; Krishnan and Raj, 2006; Benczur et al., 2005). Although making good use of link information, link-based methods neglect the content-based information. Moreover, unlike the web, there is no explicit link structure on CQA sites. So two intuitive research questions are: (1) Is there any useful link-based structure for spammer detection in CQA? (2) If so, can the two techniques, i.e., content-based model and link-based model, be integrated together to complement each other for CQA spammer detection?

To address the problems, in this paper, we first investigate the link-based structure in CQA. Then we formulate the task as an optimization problem in the graph with an efficient solution. We learn a content-based predictor as an objective function. The link-based information is incorporated into textual predictor by the way of graph regularization. Finally, to evaluate the proposed approach, we crawled a large data set from a commercial CQA site. Experimental results demonstrate that our proposed method can improve the accuracy of spammer detection.

The major contributions of this work can be summarized as follows: (1) To the best of our knowledge, our work is the first study on spammer detection on CQA sites; (2) Our proposed optimization model can integrate the advantages of both content-based model and link-based model for CQA spammer detection. (3) Experimental results demonstrate that our method can improve accuracy of spammer detection.

The remaining of the paper is organized as follows: In section 2, we review a number of the state-of-the-art approaches in related areas. Section 3 analyzes the social network of CQA sites. Section 4 presents the proposed method. Experimental results in test collections and analysis are shown in section 5. Section 6 concludes this paper.

2 Related Work

Most of current studies on spam detection can be roughly divided into two categories: content-based model and link-based model.

Content-based method targets at extracting evidences from textual descriptions of the content, treating the text corpus as a set of objects with associated attributes, and applying some classification methods to detect spam(P.Heymann et al., 2007; C.Castillo et al., 2007; Y.Liu et al., 2008; Y.Xie et al., 2008). Fetterly proposed quite a few statistical properties of web pages that could be used to detect content spam(D.Fetterly et al., 2004). Benevenuto went a step further by addressing the issue of detecting video spammers and promoters and applied the state-of-the-arts supervised classification algorithm to detect spammers and promoters(Benevenuto et al., 2009). Lee proposed and evaluated a honeypot-based approach for uncovering social spammers in online social systems(Lee et al., 2010). Wang proposed to improve spam classification on a microblogging platform(Wang, 2010).

An alternative web spam detection technique relies on link analysis algorithms, since a hyperlink often reflects some degree of similarity among pages (Gyngyi and Garcia-Molina, 2005; Gyongyi et al., 2006; Zhou et al., 2008). Corresponding algorithms include TrustRank(Z.Gyongyi et al., 2004) and AntiTrustRank(Krishnan and Raj, 2006), which used a seed set of Web pages with labels of trustiness or badness and propagate these labels through the link graph. Moreover, Benczur developed an algorithm called SpamRank which penalized suspicious pages when computing PageRank(Benczur et al., 2005).

3 Analysis on Social Network

Before analyzing the social network in CQA, we introduce some definitions. We refer users on CQA sites are someone who ask at least one question or answer at least one question. Moreover, users are divided into two categories: spammers and legitimate users. We define spammers as users who post at least one question or one answer intent to create spam.

A CQA site is particularly rich in user interactions. These interactions can be represented by Figure 1(a), where a particular question has a number of answers associated with it, represented by an edge from the question to each of the answer. We also include vertices representing authors of question or answers. An edge from a user to a question means that the user asked the question, and an edge from an answer to a user means that the answer was posted by this user. In the example,

a user U_1 asks a question Q_1, while users U_4, U_5 and U_6 answers this question. In order to observe the relation between users more clearly and directly, we summarize the relations between users as a graph shown in Figure 1(b). This graph contains vertices representing the users and omits the actual questions and answers that connect the users.

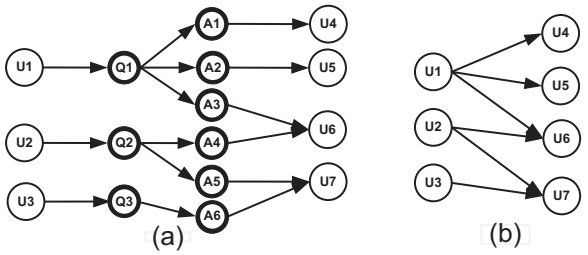

Figure 1: (a) Graph with users, questions, and answers in CQA; (b) Summary graph of users in C-QA

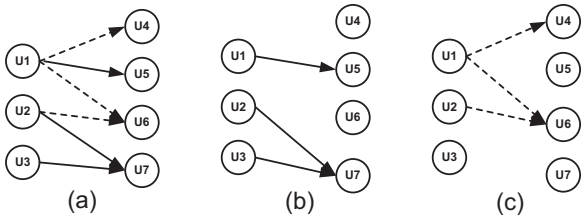

Figure 2: User graph with different relations in C-QA (a) Question-answer relation; (b) Best-answer relation; (c) Non-best-answer relation

Three kinds of major relations among users on CQA sites are defined as follows:

Question-answer relation: As shown in Figure 2(a), U_4 answers U_1's question. We define that U_4 and U_1 have Question-answer relation. Furthermore, Question-answer relation can be divided into two disjoint sets: best-answer relation and non-best-answer relation.

Best-answer relation: U_1 selects U_5's answer as the best answer. We define that U_1 and U_5 have best-answer relation. The solid lines in Figure 2(b) express the best-answer relation.

Non-best-answer relation: U_1 does not select U_4's answer as the best answer. We define that U_1 and U_4 have non-best-answer relation. The dashed lines in Figure 2(c) express the non-best-answer relation.

3.1 Best-answer Consistency Property

From analyzing data crawled from CQA site, we present the following property about best-answer relation:

Best-answer consistency property: If U_i selects U_j's answer as the best answer, the classes of users U_i and U_j should be similar.

We explain this property as follows: consider that a legitimate user is unlikely to select a spammer's answer as the best answer due to its low quality, while a legitimate user is unlikely to answer a spammer's question, so the possibility of a spammer selecting a legitimate user's answer will also be small. This means that two users linked via best-answer relation are more likely to share similar property than two random users.

3.2 Characteristics of Best Answer Spammer

Different from the general spammers, some spammers generate many fake best answers to obtain higher status in the community. We refer them as *best answer spammers*. In order to generate fake best answers, a spammer creates multiple user accounts first. Then, it uses some of the accounts to ask questions, and others to provide answers. Such spammers may post low quality answers to their own questions, and select those as the best by themselves. They may generate lots of fake best answers, which may highly impact the user experience.

Furthermore, when the spammer's intention is just advertising, we can easily identify signs of its activity: repeated phone numbers or URLs and then ignore them. However, when the spammer's intention is to obtain higher reputation within the community, the spam content may lack obvious patterns. Fortunately, there are still some clues that may help identify best answer spammers. Two characteristics are described as follows:

High best answer rate: Best answer rate is the ratio of answers selected as the best answer among the total answers. This kind of spammers have an incredible high best answer rate, compared to normal users. Specifically, in a possible best answer spammer pair, sometimes only one user has an incredible high best answer rate. Because normally one responses for asking and another for answering. So we calculate the best answer rate $BR(i,j)$ for a user pair (u_i, u_j) based on the maximum of their best answer rates:

$$BR(i,j) = Max(BR(i), BR(j)) \qquad (1)$$

Where $BR(i)$ is the best answer rate of u_i.

Time margin score: To be efficient, best answer spammers tend to answer their own ques-

tion quickly. We consider the time margin score $Time(i,j)$ between a question posted and answered for u_i and u_j as an evidence.

$$Time(i,j) = \begin{cases} 1, & \texttt{if } TimeMargin(i,j) < \varepsilon \\ 0, & \texttt{otherwise} \end{cases} \quad (2)$$

where $TimeMargin(i,j)$ is the real time margin between u_i asks a question and u_j answers this question and ε = 30 minutes.

The best answer spammer score $s(i,j)$ for a user pair (u_i, u_j) can be calculated as the combination of these two scores:

$$s(i,j) = \mu BR(i,j) + (1-\mu)Time(i,j) \quad (3)$$

μ is trade-off of two scores, here we simply set μ = 0.5. The value of $s(i,j)$ is between 0 to 1. The higher $s(i,j)$ is, the more likely u_i and u_j is a pair of the best answer spammers.

4 Spammer Detection on CQA Sites

In this section, the framework of our proposed approach is presented. First, the problem is formally defined. Next, we build a baseline supervised predictor that makes use of a variety of textual features, and then the consistency property and best answer spammer characteristics are incorporated by adding regularization to the textual predictor, last we discuss how to effectively optimize it.

4.1 Problem Statement

On CQA sites, there are three distinct types of entities: users $U = \{u_1, ...u_{l+u}\}$, answers $A = \{a_1, ...a_M\}$, and questions $Q = \{q_1, ...q_N\}$. The set of users U contains both $U_L = \{u_1, ...u_l\}$ of l labeled users and $U_U = \{u_{l+1}, ...u_{l+u}\}$ of u unlabeled users. We model the social network for U as a directed graph $G = (U, E)$ with adjacency matrix A, where $A_{ij} = 1$ if there is a link or edge from u_i to u_j and zero otherwise.

Given the input data $\{U_L, U_U, G, Q, A\}$, we want to learn a predictor c for a user u_i.

$$c(u_i) -> \{\texttt{spammer}, \texttt{legitimate user}\} \quad (4)$$

Legitimacy score y_i ($0 \leq y_i \leq 1$,i =1,2,...n) is computed for all the users. The lower y_i is, the more likely u_i is a spammer.

4.2 Text-based Spammer Prediction

In this subsection, we build a baseline predictor based on textual features in a supervised fashion. We regard the legitimacy scores as generated by combining textual features.

We consider the following textual features.

- *The Length of answers*: The length may to some extent indicate the quality of the answer. The average length of answers is calculated as a feature.
- *The ratio of Ads words in answers*: Advertising of products is the main goal of a kind of spammers and they repeat some advertisement words in their answers.
- *The ratio of Ads words in questions*: Some spammers will refer some Ads in questions in order to get attention from more users.
- *The number of received answers*: The number of received answers can indicate the quality of the question.
- *Best answer rate:* Best answer rate can show the quality of their answers.
- *The number of answers:* It can indicate the authority of a user.
- *Relevance of question and answer*: We measure the average content similarity over a pair of question and answer which is computed using the standard cosine similarity over the bag-of-words vector representation.
- *Duplication of answers*: The Jaccard similarity of answers are applied to indicate the duplication of answers .

With these features, suppose there are in total k features for each user u_i, denoted as x_i. Then $X = (x_1, x_2, ...x_n)$ is the k-by-n feature matrix of all users. Based on these features, we define the legitimacy score of each user as follows,

$$y_i = w^T x_i \quad (5)$$

where w is a k-dimensional weight vector.

Suppose we have legitimate/spammer labels t_i in the training set.

$$t_i = \begin{cases} 1, & u_i \text{ is labeled as legitimate user} \\ 0, & u_i \text{ is labeled as spammer} \end{cases} \quad (6)$$

We will then define the loss term as follows,

$$\Omega(w) = \frac{1}{l}\sum_{i=1}^{l}(w^T x_i - t_i)^2 + \alpha w^T w \quad (7)$$

Once we have learned the weight vector w, we can apply it to any user feature vector and predict the class of unlabeled users.

4.3 Regularization for Consistency Property

In Section 4.2, each user is considered as a standalone item. In this subsection, we exploit social information to improve CQA spammer detection.

In Section 3.1, the consistency property has been analyzed that users connected via best-answer relation are more similar in property. So the property is enforced by adding a regularization term into the optimization model. The regularization is acted in a collection data set, including a small amount of labeled data(l users) and a large amount of unlabeled data(u users). Then the regularization term is formulated as:

$$REG_1(U) = \sum_{i,j}^{l+u} A_{ij}(y_i - y_j)^2 \quad (8)$$

Minimizing the regularization constraint will force users who have best-answer relation belong to the same class. We formulate this as graph regularization. The graph adjacency matrix A is defined as $A_{ij} = 1$ if u_j selects u_i's answer as the best answer, and zero otherwise. Then, Equation 8 becomes:

$$REG_1(w) = \sum_{i,j}^{l+u} A_{ij}(w^T x_i - w^T x_j)^2 \quad (9)$$

With this regularization, then the objective function Equation 7 becomes:

$$\Omega_1(w) = \frac{1}{l}\sum_{i=1}^{l}(w^T x_i - t_i)^2 + \alpha w^T w + \beta \sum_{i,j}^{l+u} A_{ij}(w^T x_i - w^T x_j)^2 \quad (10)$$

4.4 Regularization for Best Answer Spammer

In this subsection, we focus on best answer spammers. Since they cannot be easily detected by only textual features(Equation 7), we introduce an additional penalty score b_i to each user u_i which indicates the possibility of becoming a best answer spammer. With the penalty score b_i, Equation 5 can be redefined as follows:

$$y_i = w^T x_i - b_i \quad (11)$$

where b_i is a non-negative score.

In order to obtain b_i, characteristics of best answer spammers are incorporated by adding graph regularization to the optimization problem. The regularization is also acted in a collection data set. Two kinds of regularization are presented as follows:

Penalty for Best Answer Spammers in Pairs

As described in Section 3.2, the score $s(i, j)$ indicates the possibility of u_i and u_j becoming a pair of best answer spammers(Equation 3). We expect u_i and u_j, who create the spam together, should share this possibility together, as follows: $b_i + b_j = e \times s(i, j)$, where e is a penalty factor, we empirically set it to 0.5.

Then we can also formulate this as graph regularization as:

$$REG_2(b) = \sum_{i<j}^{l+u} A_{ij}(b_i + b_j - e \times s(i, j))^2 \quad (12)$$

Penalty Assignment for Individual User

After introducing a penalty score to the user pair (u_i, u_j) , we have to decide how they share this penalty.

Penalty is assigned to u_i and u_j similarly. This can be also formulated as graph regularization as follows:

$$REG_3(b) = \sum_{i<j}^{l+u} A_{ij}(b_i - b_j)^2 \quad (13)$$

With the regularization for best answer spammer, the objective function becomes:

$$\Omega_3(w, b) = \frac{1}{l}\sum_{i=1}^{l}(w^T x_i - b_i - t_i)^2 + \alpha w^T w + \beta \sum_{i,j}^{l+u} A_{ij}((w^T x_i - b_i) - (w^T x_j - b_j))^2 + \gamma \sum_{i<j}^{l+u} A_{ij}(b_i + b_j - e \times s(i, j))^2 + \delta \sum_{i<j}^{l+u} A_{ij}(b_i - b_j)^2 \quad (14)$$

4.5 Optimization Problem

By considering all the components of the objective function introduced in the previous subsection, we can obtain the optimization problem. Our goal is to minimize the objective function to get optimal parameters vector w^* and penalty vector b. For solving the optimization problem, we apply a kind of limited-memory Quasi-Newton(LBFGS)(Liu and Nocedal, 1989). After obtaining the optimal parameter vector w^* and b, we can use the following scoring function $y_i = w^{*T}x_i - b_i$ to calculate scores for unlabeled users. Users with low scores will be regarded as spammers.

5 Experiments

In this section, the experimental evaluation of our approach is presented. Firstly, we introduce the details of our data sets. Then the prediction performance of our proposed approach is compared with other methods. Finally, we test the contribution of the loss term and each regularization term on these real data sets and conduct some further analysis.

5.1 Data Collections

In order to evaluate our proposed approach to detect CQA spammers from the CQA site, we need a training/test collection of users, classified into the target categories. However, to the best of our knowledge, no such collection is currently available, thus requiring us to build one.

We consider a CQA user is a user if he has posted at least one question or one answer. Moreover, we define spammer as a user who intends to create one spam. Examples of spams are: (1) an advertisement of a product or web site. (2) Completely unrelated to the subject of question. A user that is not a spammer is considered legitimate. Then we will explain the strategy of crawling data from a CQA site, Baidu Zhidao, one of the most popular CQA site in China. We randomly select 50 seed users covering different topics, including sports, entertainment, medicine and technology. The crawler follows links of question asked and question answered, gathering information on different attributes of users, including content of all responded questions and answers. The crawler ran for one week, gathering 29,257 users and 299,815 Q&A pairs. From the collection data, we randomly select a training set of 1000 users for learning process and a test set of 698 users for evaluation.

Three annotators were asked to label the users as spammers or legitimate users in both training and test set. All of the judges are Chinese and have used Baidu Zhidao frequently. The annotators judge the property of a user comprehensively based on the content information (quality of their answers, i.e. advertising and duplication of answers) and social information (interaction with other possible-spammers). The Cohen's Kappa coefficient is around 0.85, showing fair to good agreement. And our test collection contains 698 users, including 525 legitimate users and 173 spammers.

5.2 Metrics and Settings

To measure the effectiveness of our proposed method, we use the standard metrics such as precision, recall, the F1 measure. Precision is the ratio of correctly predicted users among the total predicted users by system. Recall(R) is the ratio of correctly predicted users among the actual users manually assigned. F1 is a measure that trades off precision versus recall. F1 measure of the spammer class is $2PR/(P+R)$.

We fix the parameter α in optimization method to 0.0005 which gives the best performance for the textual predictor and simply set the coefficients $\beta = 0.5$ $\gamma = \delta = 1$ in the objective function. The problem of parameter sensitivity will be tested in Section 5.6. In the optimization process, initial value of w_i is set to a random value range from 0 to 1 and initial value of b_i is set to 0.

5.3 Comparison with Other Methods

Since there has been little work on QA spam detection, we implement four state-of-the-art methods for comparison, where TrustRank and AntiTrustRank are selected to represent link-based model, while Decision Tree and SVM are two content-based classifiers.

- **Our approach**: Optimization with regularization terms that Similarity with best-answer relation, penalty for Best answer spammer. (Equation 14)

- **TrustRank**: TrustRank is a well-known link-based method in Web spam detection, which is totally based on the Web link graph(Z.Gyongyi et al., 2004).

- **AntiTrustRank**: AntiTrustRank is another well-known link-based method, which assumes that a web page pointing to spam pages is likely to be spam(Krishnan and Raj, 2006).
- **Decision Tree**: Castillo et al. applied a base classifier, decision tree, for spam detection, the features include content-based and link-based features(C.Castillo et al., 2007).
- **SVM**: We applied another state-of-the-art classifier SVM(Cortes and Vapnik, 1995). The features are the same as that used in Decision Tree method.

Methods	Precision	Recall	F1
TrustRank	0.581	0.485	0.529
AntiTrustRank	0.632	0.545	0.585
Decision Tree	0.891	0.740	0.808
SVM	0.898	0.748	0.816
Our approach	**0.925**	**0.861**	**0.892**

Table 1: Performance comparison with other methods

In Table 1, the performance of each method is listed for comparison. From the table, we have the following observations.

First, taking the advantages of both content-based model and link-based model, our optimization approach outperforms baselines under all metrics. This indicates the robustness and effectiveness of our approach.

The second observation is link-based models(**TrustRank** and **AntiTrustRank**) cannot perform well. The explanations are as follows. (1)Link-based models rely solely on hyperlinks, without considering content-based features. However, as described in section 4.2, the content can provide a strong hint for detecting spammers. (2)A technical requirement of link-based model is that the link graph must be strongly connected, which may be the case in Web, but it is not the case in QA user question-answer graph. We measured on our collection dataset and found that the graph density(defined as $D = \frac{2|E|}{|V|(|V|-1)}$ for a graph with vertices V and edges E) of user question-answer graph is only 10^{-4}. The small connectivity limits the performance of link-based model. This indicates that link-based models cannot be directly applied to CQA spammer detection. Considering that our proposed approach can integrate content-based features and link-based features effectively, we regard our approach as very complementary to the state-of-the-art link-based methods.

Another observation is that the content-based classifiers underperform our approach. And **SVM** performs slightly better than **Decision Tree**. This shows the advantages of our proposed regularization in section 4. Regularization for consistency can propagate the labeled information among users, and regularization for best answer spammers help to identify the best answer spammers.

5.4 Contribution of Loss and Regularization

In this subsection, we validate the contribution of our proposed loss term and regularization terms by the performance of real spammer detection task. And Table 2 lists the results of each method for comparison. We consider the following methods.

BL: Optimization using only content-based features. (Equation 7)

REG:Sim: Optimization with one regularization term that Similarity with best-answer relation. (Equation 10)

REG:Sim+BAS: Optimization with all regularization terms that Similarity with best-answer relation, penalty for Best Answer Spammer. (Equation 14)

Methods	Precision	Recall	F1
BL	0.911	0.711	0.798
REG:Sim	**0.945**	0.699	0.804
REG:Sim+BAS	0.925	0.861	**0.892**

Table 2: Performance of our optimization methods with different regularization for comparison

From the results we have the following observations: (1) Our content-based classifier **BL** performs well, due to the well-formed supervised learning model and reasonable features. (2) The performance of **REG:Sim** improves over **BL**, especially in the Precision measure because the social information is useful. (3) **REG:Sim+BAS** can significantly improve over **BL** especially in Recall measure. Because after adding penalty to best answer spammer, some best answer spammers can be detected successful.

5.5 Contribution of Content-based Features

In this subsection, we test the robustness of the features described in Section 4.2.

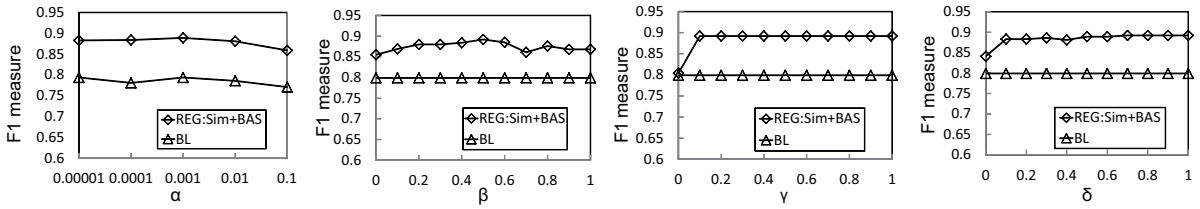

Figure 4: Parameter Sensitivity

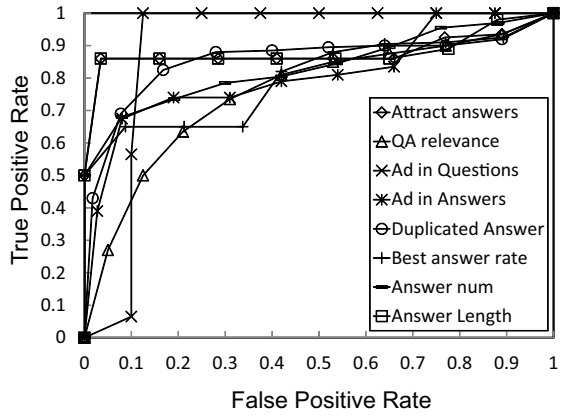

Figure 3: Content features comparison

To measure the discrimination power between spammers and legitimate users of each proposed attribute, we generate a Receiver Operating Characteristics (ROC)curve. ROC curves plot false positive rate on the X axis and true positive rate on the Y axis. The closer the ROC curve is to the upper left corner, the higher the overall accuracy is. Samples with the lowest scores (10%,20%...100%) for each attribute are labeled as spammers respectively. The (ROC) curve are shown in Figure 3. Figure 3 shows the discrimination power of each content feature we described in Section 4.2. The first observation is that all of the content features are discriminative. The feature of Ads words in questions is the most powerful. Because few legitimate users will repeat Ads words in questions, so this feature can help to identify spammers more easily. Note that the feature of the best answer rate do not perform well. Because some best answer spammers also have high best answer rate.

5.6 Parameter Sensitivity

Our optimization approach have four parameters $\alpha, \beta, \gamma, \delta$ to set: the tradeoff weight for each regularization term. The value of the regularization weight controls our importance in the regularizer: a higher value results in a higher penalty when violating the corresponding regularization. So we mainly evaluate the sensitivity of our model with parameters by fixing all the other parameters and let one of $\{\alpha, \beta, \gamma, \delta\}$ varies. Figure 4 shows the prediction performance in F1 measure varying each parameter. As we observed over a large range of parameters, our approach (**REG:Sim+BAS**) achieves significantly better performance than **BL** method. It indicates that the parameters selection will not critically affect the performance of our optimization approach.

6 Conclusion

In this paper, we first studied social networks on CQA sites. We found that spammers are usually connected to other spammers via the best-answer relation. We also studied the "best answer spammers" on CQA sites, which cannot be easily detected for lack of identifiable textual patterns. Our proposed model incorporated the link-based information by adding regularization constraints to the textual predictor. Experimental results demonstrated that our method is more effective for spammer detection compared to other state-of-the-art methods. Besides obtaining better performance, we have also analyzed the CQA social networks, which gives us insight on the model design.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (61003092, 61073069), National Major Science and Technology Special Project of China (2014ZX03006005), Shanghai Municipal Science and Technology Commission (No.12511504500) and "Chen Guang" project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation(11CG05).

References

Andras A. Benczur, Karoly Csalogany, Tamas Sarlos, and Mate Uher. 2005. Spamrank-fully automatic link spam detection. In *AIRWeb'05*.

Fabricio Benevenuto, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida, and Marcos Goncalves. 2009. Detecting spammers and content promoters in online video social networks. In *Proceeding of SIGIR*.

Maarten Bosma, Edgar Meij, and Wouter Weerkamp. 2012. A framework for unsupervised spam detection in social networking sites. In *Proceedings of ECIR*.

C.Castillo, D.Donato, A.Gionis, V.Murdock, and F.Silvestri. 2007. Know your neighbors: Web spam detection using the web topoloty. In *Int'l ACM SIGIR*.

Zhicong Cheng, Bin Gao, Congkai Sun, Yanbing Jiang, and Tie-Yan Liu. 2011. Let web spammers expose themselves. In *WSDM*.

Corinna Cortes and Vlandimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20:273–297.

C.Wu, K.Cheng, Q.Zhu, and Y.Wu. 2005. Using visual features for anti-spam filtering. In *IEEE Int'l Conference on Image Processing(ICIP)*.

D.Fetterly, M.Manasse, and M.Najork. 2004. Spam,damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Int'l Workshop on the Web and Databases(WebDB)*.

Zoltn Gyngyi and Hector Garcia-Molina. 2005. Link spam alliances. In *VLDB*.

Zoltan Gyongyi and Hector Garcia Molina. 2004. Web spam taxonomy. Technical report, Stanford Digital Library Technologies Project.

Zoltan Gyongyi, PavelBerkhin, Hetor Garcia-Molina, and Jan O. Pedersen. 2006. Link spam detection based on mass estimation. In *VLDB*.

Vijay Krishnan and Rashmi Raj. 2006. Web spam detection with anti-trust rank. In *ACM SIGIR workshop on adversarial information retrieval on the Web*.

Kyumin Lee, James Caverlee, and Steve Webb. 2010. Uncovering social spammers: Social honeypots + machine learning. In *Proceeding of SIGIR*.

L.Gomes, J.Almeida, V.Almeida, and W.Meira. 2007. Workload models of spam and legitimate e-mails. In *Performance Evaluation*.

Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady W. Lauw. 2010. Detecting product review spammers using rating behaviors. In *CIKM*.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45:503–528.

N.Jindal and B.Liu. 2008. Opinion spam and analysis. In *WSDM*.

Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of WWW*.

Myle ott, Yejin Choi, Claire Cardie, and Jeffrey T.Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *ACL*.

P.Heymann, G.Koutrika, and H.Garcia-Molina. 2007. Fighting spam on social web sites: A survey of approaches and future challenges. In *IEEE Internet Computing*.

Xiance Si, Edward Y. Chang, Zoltan Gyongyi, and Maosong Sun. 2010. Confucius and its intelligent disciples: integrating social with search. In *Proceeding of VLDB*.

Alex Hai Wang. 2010. Don't follow me: Twitter spam detection. In *Proceedings of 5th International Conference on Security and Cryptography (SECRYPT)*.

Y.Liu, H.Sundaram, Y.Chi, J.Tatemura, and B.Tseng. 2008. Detecting splogs via temporal dynamics using self-similarilty analysis. In *ACM Transactions on the Web(TWeb)*.

Y.Xie, F.Yu, K.Achan R.Panigrahy, G.Hulten, and I.Osipkov. 2008. Spamming botnet: Signatures and characteristics. In *ACM SIGCOMM*.

Z.Gyongyi, H.Garcia-Molina, and J.pedersen. 2004. Combating web spam with trustrank. In *Int'l Conference on Very Large Data Bases(VLDB)*.

Bin Zhou, Jian Pei, and ZhaoHui Tang. 2008. A spamicity approach to web spam detection. In *SDM*.

Yin Zhu, Xiao Wang, Erheng Zhong, Nanthan N. Liu, He Li, and Qiang Yang. 2012. Discovering spammers in social networks. In *Proceedings of AAAI*.