Detecting Promotion Campaigns in Community Question Answering

Xin Li[†], Yiqun Liu[†], Min Zhang[†], Shaoping Ma[†], Xuan Zhu[‡], Jiashen Sun[‡]
[†]State Key Lab of Intelligent Technology & Systems; Tsinghua National TNLIST Lab
[†]Department of Computer Science & Technology, Tsinghua University, Beijing, 100084, China
[‡]Samsung R&D Institute China - Beijing
x-l08@163.com, {yiqunliu,z-m,msp}@tsinghua.edu.cn, {xuan.zhu,jiashens.sun}@samsung.com

Abstract

With Community Question Answering (CQA) evolving into a quite popular method for information seeking and providing, it also becomes a target for spammers to disseminate promotion campaigns. Although there are a number of quality estimation efforts on the CQA platform, most of these works focus on identifying and reducing low-quality answers, which are mostly generated by impatient or inexperienced answerers. However, a large number of promotion answers appear to provide high-quality information to cheat CQA users in future interactions. Therefore, most existing quality estimation works in CQA may fail to detect these specially designed answers or question-answer pairs. In contrast to these works, we focus on the promotion channels of spammers, which include (shortened) URLs, telephone numbers and social media accounts. Spammers rely on these channels to connect to users to achieve promotion goals so they are irreplaceable for spamming activities. We propose a propagation algorithm to diffuse promotion intents on an "answerer-channel" bipartite graph and detect possible spamming activities. A supervised learning framework is also proposed to identify whether a QA pair is spam based on propagated promotion intents. Experimental results based on more than 6 million entries from a popular Chinese CQA portal show that our approach outperforms a number of existing quality estimation methods for detecting promotion campaigns on both the answer level and QA pair level.

1 Introduction

In the last decade, Community Question Answering (CQA) portals have emerged as a popular platform for individuals to seek and provide information. Because of the large number of users, CQA portals have accumulated a tremendous number of questions and answers [Wu *et al.*, 2014; Ji and Wang, 2013; Zhou *et al.*, 2012]. However, a relatively high proportion of answers in CQA are of low quality [Agichtein *et al.*, 2008; Sakai *et al.*, 2011]. Spammers also expose promotion campaigns to CQA users to advance their commercial interests [Ding *et al.*, 2013]. Additionally, some crowdsourcing systems such as Zhubajie[1] provide paid services to organize promotion campaigns on CQA portals [Chen *et al.*, 2013; Wang *et al.*, 2014; Tian *et al.*, 2015] and try to gain profit by attracting users to certain Web sites or by persuading them to buy certain products.

Although fraudulent information and spam information are usually contained in promotion campaigns, which makes the CQA environment less credible and more noisy, few techniques exist to help CQA portals identify and warn users of these promotion activities. Most existing works focus on estimating the quality of answers or question-answer (QA) pairs [Liu *et al.*, 2008; Agichtein *et al.*, 2009; Liu *et al.*, 2011] in a CQA environment. However, on one hand, low-quality answers do not necessarily contain promotion information because the answerer may only be unfamiliar with the question or inexperienced with the interaction process in CQA. On the other hand, answers containing promotion information are not necessarily low quality according to the traditional quality assessment standards because spammers may organize their answers sufficiently to make it more appealing to users. Paid experts from crowdsourcing systems may even carefully design the QA pairs to make them similar to legitimate QA pairs. Table 1 shows an example extracted from a popular Chinese CQA portal that contains promotion information[2]. The first part of the answer contains some high-quality suggestions for the questioner. While at the end of the answer, a product is promoted by the answerer via a shortened URL, which makes it part of a promotion campaign. This answer may be identified as a high-quality answer by existing quality estimation methods because it actually provides some useful information. However, the promotion information provided along with the useful information may be misleading (in this special case, the product is illegal and banned for side effects).

Our work differs from existing efforts in detecting self-answer spamming posts in CQA portals such as [Chen *et al.*, 2013]. These works aim to address a special case of promotion campaign in which spammers ask questions and select their self-posted answers (usually with spamming information) as best answers to attract CQA users. However, there

[1]http://www.zhubajie.com/
[2]http://wenwen.sogou.com/z/q581397464.htm

Table 1: Example of a QA pair with promotion information from a Chinese CQA portal (contents are translated into English)

Question	How do I get rid of body odor?
Answer	It does not matter. Anybody can have some sort of body odor. You need to pay attention to personal hygiene and take showers frequently. Keep a regular lifestyle and a stable mood. You may also try this product http://t.cn/RvGjjvg. I recovered from body odor by using it.

are many occasions where spammers answer existing questions from legitimate users with seemingly high quality contents as shown in Table 1. Their answers may be selected by questioners as best answers because of the high-quality information provided. Even if they are not selected as best answers, they will also be displayed along with the question and attract possible interactions. Consequently, we need to detect promotion campaigns in both the self-answer scenario and other scenarios.

In this paper, we propose a framework to detect promotion campaigns both on the answer level and QA pair level. Based on the assumption that 1. spammers will use promotion channels (such as (shortened) URLs, telephone numbers and social media accounts) to organize promotion campaigns and that 2. spammers usually use one CQA account to promote multiple products and one product usually relies on multiple accounts, we propose a detection algorithm based on the propagation of the promotion intents of spammers. On the answer level, we start by selecting a small set of promotion channels and construct an "answerer-channel" bipartite graph based on the channels that answerers post. After that, we propose a propagation algorithm to diffuse the spamming scores of seed promotion channels on the bipartite graph to detect additional spamming channels and spammers. On the QA pair level, we introduce the spamming scores of users and promotion channels calculated with the propagation algorithm as features and apply a supervised learning model to identify whether a QA pair is spam.

The main contributions of this paper can be summarized as follows:

- To the best of our knowledge, this is the first attempt to identify promotion campaigns on CQA portals both on the answer level and QA pair level.
- An approach to identify promotion campaigns is proposed based on the close relationships between spammers and promotion channels.
- An evaluation dataset is constructed that contains millions of entries from a popular Chinese CQA portal and a large number of annotated fraudulent/legitimate answers and QA pairs.

The remainder of this paper is organized as follows: after a discussion of the related work in the next section, we discuss the existence of different promotion channels in the third section. Then, we introduce the promotion campaign identification algorithm in the fourth section. The fifth section presents an evaluation and discussion of our approach. Finally, the sixth section concludes the paper.

2 Related Work

Most existing spam detection approaches on CQA portals focus on estimating the quality of answers or QA pairs. [Harper *et al.*, 2008] investigates the predictors of answer quality and find that answer quality is typically higher on fee-based sites versus free sites, and higher pay for answers usually leads to better outcomes. [Suryanto *et al.*, 2009] proposes a quality-aware framework to retrieve answers from a CQA portal based on both answer content and the expertise of answerers. Experimental results show that expertise based methods outperform methods using answer content features only. In addition to estimating answer qualities, [Li *et al.*, 2012] estimate question quality with a Mutual Reinforcement-based Label Propagation algorithm. Despite the success of these methods in detecting low-quality answers and QA pairs, we cannot equate low-quality answers (QA pairs) with answers (QA pairs) that contain promotion campaigns (e.g., the instance in Table 1). Therefore, they may not be suitable for the task of identifying promotion campaigns.

[Chen *et al.*, 2013] propose a method to detect commercial campaigns in best answers from the CQA portals. They argue that widely used features such as textual similarities between questions and answers will no longer be effective to filter commercial paid posters. Therefore, they combine more context information, such as writing templates and a user's reputation track, to form a new model to detect the potential campaign answers. Their detection method integrates semantic analysis and poster track records and utilizes the special features of CQA websites, which shows great potential towards adaptive online detection performance. However, it is a special case of promotion campaigns where paid posters ask questions and select their self-posted answers as best answers. We also need to detect more general cases where spammers answer existing questions from legitimate users and use promotion campaigns in their answers, which are not considered in existing works.

3 Promotion Channels

CQA users generally seek instant solutions when they post questions in the community. They prefer short and targeted answers to their questions. As a result, the information contained in the answers is limited. We believe that an answer itself usually cannot cheat users. Instead, spammers rely on some channels to link users and their promotion goals, which are irreplaceable for spamming activities. In this paper, we focus on three types of promotion channels: (shortened) URL, telephone number and social media account based on our observation of CQA portals.

- **URL**: (shortened) URL is the most widely used channel by spammers to promote their products (see Table 1 for an example). The URL will lead to an e-commerce website, which usually shows the description of a product or even links to make a purchase.
- **Telephone number**: fraudulent telephone numbers are often used by spammers to cheat Web users [Li *et al.*,

2014a]. On CQA portals, spammers inject telephone numbers into their answers and attract users to dial. Then, they will persuade the users to buy their products/services or obtain the user's personal information for illegal purposes.

- **Social media account**: social media such as QQ and WeChat[3] provide spammers with a new method to perform promotion campaigns. Spammers leave their social media accounts in the answers. When CQA users communicate with them via social media, the spammers will be able to conduct their spamming activities.

Tables 2 and 3 show two example QA pairs containing a telephone number[4] and social media account[5], separately.

Table 2: Example of a QA pair containing telephone number

Question	What business gift should I send?
Answer	Xxx company produces specially designed gifts, including business gifts, conference gifts, birthday gifts, etc. Please contact 15549083151.

Table 3: Example of a QA pair containing QQ account

Question	Which brand of acne product is better?
Answer	You can try xxx mask, which is good for removing acnes. Contains no hormones. If you need assistance, please contact qq 252045995.

4 Promotion Campaign Detection in CQA

In this section, we introduce the framework of our approach. Its flowchart is shown in Figure 1, which is mainly composed of two parts: answer level promotion campaign detection and QA pair level promotion campaign detection.

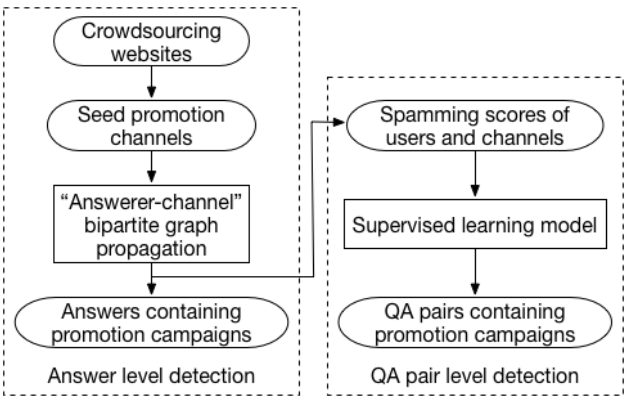

Figure 1: Flowchart of our approach

[3]QQ and WeChat are the two most popular social platforms in China.

[4]http://wenwen.sogou.com/z/q507423589.htm

[5]http://wenwen.sogou.com/z/q574609598.htm

4.1 Answer level promotion campaign detection

For answer level detection of promotion campaigns, we first select a small set of seed promotion channels from a crowdsourcing website. Then, we construct an "answerer-channel" bipartite graph based on users who include promotion channel information in their answers. Finally, we propose a propagation algorithm to diffuse the spamming scores of seed promotion channels on the bipartite graph to detect additional answers that contain promotion campaigns.

Selecting seed promotion channels

Previous studies show that malicious crowdsourcing systems have been rapidly growing in both user base and total revenue [Wang *et al.*, 2012]. Take Zhubajie, a popular crowdsourcing website in China, as an example. It provides paid services for organizing promotion campaigns on various social media websites, such as microblogging, web forums, and CQA portals. In the case of CQA portals, a company can post a request on the website, which may contain the description of the products, promotion channels and the preferred CQA portals. After that, service providers may find the request and accept it. Alternatively, the company may search for service providers that can organize promotion campaigns on CQA portals and select a preferred one for service. After the company makes a payment, the paid posters will proceed with the request. They may answer existing questions from legitimate users and provide promotion information in their answers. Alternatively, they may create multiple accounts, use some of them to ask questions and use others to answer their own questions and inject promotion information. They may even select their self-posted answers as best answers to attract more CQA users. After the transaction completes similar to e-commerce websites, the merchant can assess and rate the service provider.

We aim to select a set of promotion channels from the transactions. However, the details of the transactions are hidden by the website. Instead, we can visit the homepages of users. On a user's homepage, we can see the requests that the user has posted and the service providers that have completed each of the requests. We visit the homepages of 10,000 users and extract promotion channels from the descriptions of requests that are completed by CQA service providers. As a result, we obtain a set of promotion channels for 106 URLs, 15 telephone numbers, 19 QQ accounts and eight WeChat accounts because a large portion of users never posted requests to promote products on CQA portals. The selection of promotion channels from the crowdsourcing website is quite precise (with 100% precision) because they are extracted from the requests to promote products on CQA portals. Hence, they can be used as seeds to find more promotion channels and answers that contain promotion campaigns.

Constructing "answerer-channel" bipartite graph

On CQA portals, a user may provide various answers to different questions. Meanwhile, a promotion channel may be involved in different answers. Figure 2(a) shows the relations of answerers, answers and promotion channels. If we only reserve the answerer side and promotion channel side as shown in Figure 2(b), we can obtain a simplified graph between answerers and promotion channels, thus constructing

the "answerer-channel" bipartite graph. Here, we define the weight matrix of the bipartite graph $W = \{w_{ij}\}_{n \times m}$, where n is the total number of users in CQA, m is the total number of promotion channels, and w_{ij} is the frequency of user i involving promotion channel j in all of the answers the user has posted.

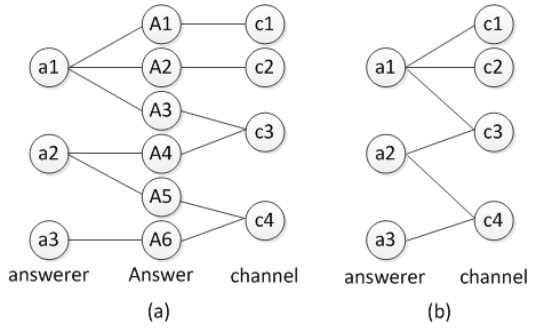

Figure 2: (a) Relation graph of answerers, answers and promotion channels; (b) Simplified relation graph between answerers and promotion channels

Propagation on the "answerer-channel" bipartite graph
After collecting a small set of seed promotion channels and constructing the "answerer-channel" bipartite graph, our goal is to diffuse the promotion intents of the seed promotion channels on the bipartite graph and detect possible spam answers. So, we propose an "answerer-channel" bipartite graph propagation algorithm, which is based on the following assumptions:

Assumption 1 (Channel assumption) *Spammers need to use promotion channels to organize promotion campaigns.*

We believe that an answer itself does not contain sufficient details to attract users into further interactions. Instead, spammers rely on promotion channels to achieve promotion goals, which make them irreplaceable for spamming activities.

Assumption 2 (Concurrence assumption) *If a certain channel contained in a certain user's answers proves to be a promotion channel, the other channels contained in the user's answers are likely to be involved in promotion campaigns as well.*

Usually, spammers gain profit by organizing multiple promotion campaigns. Because a certain promotion campaign is related to only one channel or a few channels, it is reasonable to assume that the other channels posted by the spammer are also promotion campaigns (although they may not be involved in the same campaign).

Based on the assumptions, we can diffuse the spamming scores of seed promotion channels on the "answerer-channel" bipartite graph. We adopt a similar algorithm to [Li *et al.*, 2014b], which has proved to be effective on label diffusion. The description of the algorithm is shown in Algorithm 1.

After the iteration completes, each user and each channel will be assigned a spamming score. Because the initial score of the seed promotion channels is one, a higher score indicates a larger likelihood that the user or the channel is involved in propagation campaigns. The evaluation of answer level promotion campaign detection will be shown in Section 5.2.

Algorithm 1 Answerer-channel bipartite graph propagation algorithm

Require:
The set of seed promotion channels, S;
The set of users, U;
The set of channels, C;
The answerer-channel weight matrix, W;
The threshold to end the iteration, ϵ;

for each c_j in C **do**
  $cscore(c_j) = 0$
end for
for each c_j in S **do**
  $cscore(c_j) = 1$
end for
$n = 1$
while $|cscore_n - cscore_{n-1}| > \epsilon$ **do**
  for each u_i in U **do**
    $uscore(u_i) = \sum_i w_{ij} \times cscore(c_j)$
  end for
  for each c_j in C **do**
    if c_j in S **then**
      $cscore(c_j) = 1$
    else
      $cscore(c_j) = \sum_j w_{ij} \times uscore(u_i)$
    end if
  end for
  $n = n + 1$
end while

4.2 QA pair level promotion campaign detection

In CQA portals, a user can be both a questioner and an answerer at the same time. As mentioned above, spammers may ask questions and select their self-posted answers as best answers. As a result, we need to take into account both aspects of questioners and answerers when detecting promotion campaigns on the QA pair level. With the answer level detection method, we have obtained the spamming scores of each user and each channel. With these scores, we can adopt a supervised learning model to decide whether a QA pair belongs to a promotion campaign. Three features are extracted for each QA pair in this detection process, which are described below:

- The spamming score of the questioner.
- The spamming score of the answerer.
- The highest spamming score among all channels in the answer content.

The main difference between our approach and previous spam detection methods on CQA portals is that we consider three aspects of information: the questioner, the answerer and the promotion channel in the answer. We construct our model with logistic regression (most of the previous quality estimation methods on CQA portals use the logistic regression

model; we also tried other classification models, but the logistic regression model has the best performance). To show the effectiveness of the proposed method, we compare its performance with a number of existing quality estimation methods for CQA. We also add the proposed features to the feature set of those methods to see if they can improve the performances of the original methods. The evaluation of QA pair level promotion campaign detection will be shown in Section 5.3.

5 Experimental Results and Discussions

5.1 CQA Dataset

With the help of a popular Chinese CQA portal named Sogou Wenwen (http://wenwen.sogou.com/), we collect 6,452,981 entries (here an entry is defined as a Web page with a question and all its corresponding answers) and 11,758,802 answers with a random sampling strategy. The statistics of possible promotion channels in this data set is shown in Table 4.

Table 4: The number and proportion of entries and answers that contain promotion channels. Here, "containing a channel" means the channel is contained in at least one answer of the entry.

	Entry	Answer
URL	291,304 (4.5%)	326,576 (2.8%)
Telephone number	37,662 (0.6%)	43,550 (0.4%)
QQ account	52,657 (0.8%)	60,960 (0.5%)
WeChat account	18,840 (0.3%)	23,277 (0.2%)

We found that co-occurrences of promotion channels in the same answers are rather rare and only cover approximately 1% of the answers. As a result, a total of approximately 450,000 answers contain at least one promotion channel, involving approximately 63,000 answerers. We construct the "answerer-channel" bipartite graph with the extracted answerers, promotion channels and the posting relationships between them.

5.2 Performance of answer level detection

We diffuse the spamming scores of seed promotion channels on the bipartite graph. As stated in Section 4.1, we have obtained a seed set of 106 URLs, 15 telephone numbers, 19 QQ accounts and 8 WeChat accounts from the crowdsourcing promotion campaign service websites. To evaluate the contribution of different types of promotion channels, we incrementally add each type of seed promotion channels into the bipartite graph. Each time a new type of seed promotion channel is added into the bipartite graph, we run Algorithm 1 and see how it performs. We believe that spammers on CQA portals may answer questions without posting promotion campaigns to pretend to be legitimate users, while legitimate users will not provide promotion channels in their answers. So, the spamming score of a promotion channel is more representative of the possibility of an answer containing promotion campaigns. As a result, we define the spamming score of an answer as the highest spamming score of all of the promotion channels in it. We calculate the cumulative distribution function (CDF) for the spamming scores of answers each time we add a new type of seed promotion channel, and the comparison of CDFs is shown in Figure 3.

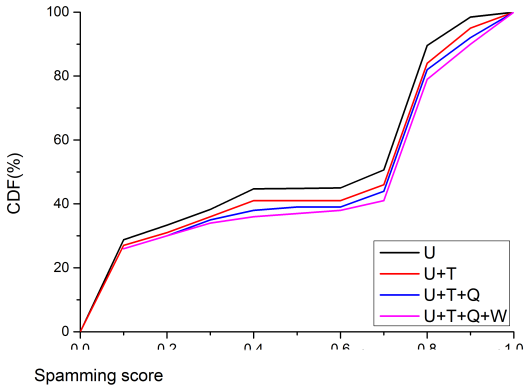

Figure 3: CDFs for spamming scores of answers diffused from a different set of seed promotion channels (U: URL, T: Telephone number, Q: QQ account, W: WeChat account)

As shown in Figure 3, with the increase of seed promotion channels, more answers receive a relatively high spamming score. If we only use the URL as the seed set, only 2% of the answers receive scores higher than 0.9. Meanwhile, if we use the whole set of seed promotion channels, the percentage of answers with scores over 0.9 is approximately 10%. Therefore, incorporating different types of promotion channels actually helps us find more possible spamming answers.

To compare the accuracy of promotion campaign detection from different sets of seed promotion channels, we randomly sample 500 answers and manually label them as spam or nonspam. If the answer contains promotion campaign information, we label it as spam. Otherwise, it is labeled as nonspam. Each time we run the algorithm with a different set of seed promotion channels, the answers receive different scores, thus having different rankings. We rank the 500 answers in descending order of the spamming score each time the algorithm completes and compares the ROC curves and the corresponding Area Under Curve (AUC) values. The results are shown in Figure 4 and Table 5.

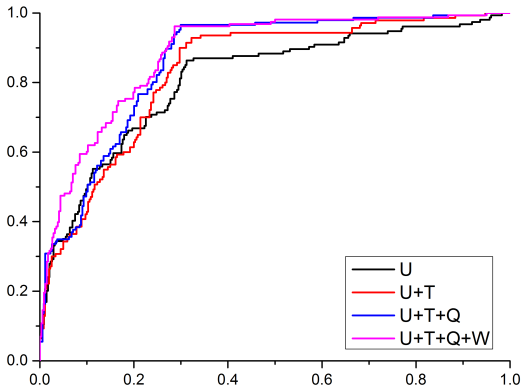

Figure 4: Comparison of ROC curves for detection methods from different sets of seed promotion channels

Table 5: Comparison of AUC values for detection methods from different sets of seed promotion channels

Seed promotion channel	AUC value	Improvement
U	0.8095	–
U+T	0.8345	3.1%
U+T+Q	0.8634	6.7%
U+T+Q+W	**0.8839**	**9.2%**

From the results, we can see that the increase in seed promotion channels contributes to the performance of promotion campaign detection in answers. Because seed promotion channels add more information into the bipartite graph, the increase in seed promotion channels helps to detect more answers containing promotion campaigns and rank them higher, thus improving the AUC value.

5.3 Performance of QA pair level detection

In QA pair level promotion campaign detection, we use the spamming scores of users and channels diffused from the whole set of seed promotion channels. To evaluate the performance of our approach, we use several QA pair quality estimation methods proposed in existing works as baselines. These methods extract content-based, behavior-based or structure-based features from QA pairs as shown below (some of the features cannot be obtained from our dataset, so we only reserve the following features):

- **Baseline1 [Jeon *et al.*, 2006]**: answerer's acceptance ratio, answer length, questioner's self-evaluation, answerer's activity level, answerer's category specialty, click counts, number of answers.
- **Baseline2 [Shah and Pomerantz, 2010]**: length of the question's subject, length of the question's content, number of answers for the question, number of comments for the question, information from the asker's profile, length of the answer's content, reciprocal rank of the answer in the list of answers for the given question, information from the answerer's profile.
- **Baseline3 [Chen *et al.*, 2013]**: interval post time, number of other answers, relevance between the questions and the answers.
- **Baseline4 [Zhang *et al.*, 2014]**: answer length, fraction of best answers an answerer is awarded in all answers he or she provides, unique number of words in an answer, word overlap between a question and an answer.

To obtain the training data, we locate the questions for each of the 500 labeled answers as described in Section 5.1 and label the 500 QA pairs with three levels:

- *Explicit Spam*: the answer contains promotion campaign information and does not meet the questioner's information needs.
- *Implicit spam*: the answer contains promotion campaign information but somehow meets the questioner's information needs.
- *Non-spam*: the answer does not contain any promotion campaign information.

We compare the performance of the three features (spamming score of the questioner, the answerer, and the highest spamming score among all channels) obtained from the propagation algorithm as stated in Section 4.2 (denoted as Promotion) with the baseline features. We also add the three features into the baseline feature set to see if the performance can be improved. In the first experiment, we only regard the explicit spam label as positive. While in the second experiment, we regard both explicit spam and implicit spam labels as positive. In each of the experiments, we apply a logistic regression model and use 10-fold cross validation. The results are shown in Table 6 and Table 7.

Table 6: Comparison of F1 scores when only regarding explicit spam label as positive

	Original baseline	With Promotion
Promotion	–	0.812
Baseline1	0.798	0.829(+3.9%)
Baseline2	0.817	0.835(+2.2%)
Baseline3	0.774	0.821(+6.1%)
Baseline4	0.752	0.815(+8.4%)

Table 7: Comparison of F1 scores when regarding both explicit spam and implicit spam labels as positive

	Original baseline	With Promotion
Promotion	–	0.819
Baseline1	0.722	0.849(+17.6%)
Baseline2	0.747	0.861(+15.3%)
Baseline3	0.701	0.838(+19.5%)
Baseline4	0.556	0.812(+46.0%)

From the results, we can see that if we only regard the explicit spam label as positive, the proposed method based on promotion intent propagation can achieve a comparable or slightly better performance compared with baseline methods. Because explicit spam QA pairs are also of low quality according to the definition, traditional quality estimation methods can usually detect them successfully. We can also find that baseline methods with the promotion features always perform better than the original solutions. Meanwhile, when we regard both explicit spam and implicit spam labels as positive, the propagation features achieve a considerably better performance than baseline methods. Additionally, after adding the propagation features into the baseline feature sets, the F1 scores improve greatly. The results confirm our assumption that implicit spam QA pairs are of high quality and cannot be detected by traditional quality estimation methods. The proposed algorithm, however, still works well on these cases and helps existing methods detect spam.

6 Conclusion

In this paper, we propose a framework to detect promotion campaigns in CQA both on an answer level and on a QA pair level. Experimental results show that different types of promotion channels can all contribute to the detection of answers containing promotion information. Moreover, spamming scores diffused from the propagation algorithm can be adopted to effectively detect both explicit and implicit spam QA pairs.

7 Acknowledgments

This work was supported by National Key Basic Research Program (2015CB358700) and Natural Science Foundation (61472206, 61073071) of China. Part of the work has been done at the Tsinghua-NUS NExT Search Centre, which is supported by the Singapore National Research Foundation & Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490).

References

[Agichtein *et al.*, 2008] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *WSDM*, pages 183–194. ACM, 2008.

[Agichtein *et al.*, 2009] Eugene Agichtein, Yandong Liu, and Jiang Bian. Modeling information-seeker satisfaction in community question answering. *TKDD*, 3(2):10, 2009.

[Chen *et al.*, 2013] Cheng Chen, Kui Wu, Venkatesh Srinivasan, and Kesav Bharadwaj R. The best answers? think twice: Online detection of commercial campaigns in the cqa forums. In *ASONAM*, pages 458–465. IEEE, 2013.

[Ding *et al.*, 2013] Zhuoye Ding, Yeyun Gong, Yaqian Zhou, Qi Zhang, and Xuanjing Huang. Detecting spammers in community question answering. In *IJCNLP*, pages 118–126, 2013.

[Harper *et al.*, 2008] F Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A Konstan. Predictors of answer quality in online q&a sites. In *SIGCHI*, pages 865–874. ACM, 2008.

[Jeon *et al.*, 2006] Jiwoon Jeon, W Bruce Croft, Joon Ho Lee, and Soyeon Park. A framework to predict the quality of answers with non-textual features. In *SIGIR*, pages 228–235. ACM, 2006.

[Ji and Wang, 2013] Zongcheng Ji and Bin Wang. Learning to rank for question routing in community question answering. In *CIKM*, pages 2363–2368. ACM, 2013.

[Li *et al.*, 2012] Baichuan Li, Tan Jin, Michael R Lyu, Irwin King, and Barley Mak. Analyzing and predicting question quality in community question answering services. In *WWW*, pages 775–782. ACM, 2012.

[Li *et al.*, 2014a] Xin Li, Yiqun Liu, Min Zhang, and Shaoping Ma. Fraudulent support telephone number identification based on co-occurrence information on the web. In *AAAI*, pages 108–114. AAAI Press, 2014.

[Li *et al.*, 2014b] Xin Li, Min Zhang, Yiqun Liu, Shaoping Ma, Yijiang Jin, and Liyun Ru. Search engine click spam detection based on bipartite graph propagation. In *WSDM*, pages 93–102. ACM, 2014.

[Liu *et al.*, 2008] Yandong Liu, Jiang Bian, and Eugene Agichtein. Predicting information seeker satisfaction in community question answering. In *SIGIR*, pages 483–490. ACM, 2008.

[Liu *et al.*, 2011] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Evgeniy Gabrilovich, Yoelle Maarek, Dan Pelleg, and Idan Szpektor. Predicting web searcher satisfaction with existing community-based answers. In *SIGIR*, pages 415–424. ACM, 2011.

[Sakai *et al.*, 2011] Tetsuya Sakai, Daisuke Ishikawa, Noriko Kando, Yohei Seki, Kazuko Kuriyama, and Chin-Yew Lin. Using graded-relevance metrics for evaluating community qa answer selection. In *WSDM*, pages 187–196. ACM, 2011.

[Shah and Pomerantz, 2010] Chirag Shah and Jefferey Pomerantz. Evaluating and predicting answer quality in community qa. In *SIGIR*, pages 411–418. ACM, 2010.

[Suryanto *et al.*, 2009] Maggy Anastasia Suryanto, Ee Peng Lim, Aixin Sun, and Roger HL Chiang. Quality-aware collaborative question answering: methods and evaluation. In *WSDM*, pages 142–151. ACM, 2009.

[Tian *et al.*, 2015] Tian Tian, Jun Zhu, Fen Xia, Xin Zhuang, and Tong Zhang. Crowd fraud detection in internet advertising. In *WWW*. ACM, 2015.

[Wang *et al.*, 2012] Gang Wang, Christo Wilson, Xiaohan Zhao, Yibo Zhu, Manish Mohanlal, Haitao Zheng, and Ben Y Zhao. Serf and turf: crowdturfing for fun and profit. In *WWW*, pages 679–688. ACM, 2012.

[Wang *et al.*, 2014] Gang Wang, Tianyi Wang, Haitao Zheng, and Ben Y Zhao. Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers. In *23rd USENIX Security Symposium, USENIX Association, CA*, 2014.

[Wu *et al.*, 2014] Haocheng Wu, Wei Wu, Ming Zhou, Enhong Chen, Lei Duan, and Heung-Yeung Shum. Improving search relevance for short queries in community question answering. In *WSDM*, pages 43–52. ACM, 2014.

[Zhang *et al.*, 2014] Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou. Question retrieval with high quality answers in community question answering. In *CIKM*, pages 371–380. ACM, 2014.

[Zhou *et al.*, 2012] Guangyou Zhou, Kang Liu, and Jun Zhao. Joint relevance and answer quality learning for question routing in community qa. In *CIKM*, pages 1492–1496. ACM, 2012.