Propagating Anti-TrustRank with Relationship Strength for Fighting Link Farming on Twitter

Hua SHEN[1,2]
[1] College of Mathematics and Information Science
Anshan Normal University
[2] School of Software Dalian University of Technology
Anshan Liaoning, China
shenhua_as@126.com

Xinyue LIU
School of Software
Dalian University of Technology
Dalian Liaoning,China
xyliu@dlut.edu.cn

***Abstract*—How to find the features of spammers and link structure is a fundamental issue in social networks. Existing work solves this problem by classification and ranking. However, these methods ignore relationship strength on edges, i.e., the weight between a pair of users. Then the challenge is how to effectively estimate the relationship strength given a pair of users. To tackle this challenge, the paper proposes a Poisson regression-based latent variable model to estimate relationship strength by jointly modelling users' similarities and interactions' frequency. Furthermore, Anti-Trust Rank with relationship strength algorithm is proposed to fight link farming on Twitter. Experimental results show that the proposed scheme can demote spammers and penalize uses that link to spammers effectively.**

Keywords-Anti-Trustrank; Social Spam; Relationship Strength

I. INTRODUCTION

In recent years, Twitter has emerged as a popular online social network for people to share or discover real-time information. Millions of users publish information about their life freely and discover topics they are interested in immediately. Unfortunately, Twitter has also attracted the attention of social spammers, who strive to achieve their malicious goals with various spam strategies[1]. Link farming is a sophisticated strategy, which is the process that spammers exchange of their links to gain influence in social networks [2]. By constructing link farming, spammers not only enhance their social influence, but infiltrate into the Twitter network to evade spammer detection. More seriously, most of farmed links come from a number of normal Twitter users due to social etiquette [2, 3], which causes great difficulty to fight social spammers effectively. Therefore, it is a challenging work to combat link farming on Twitter.

Link farming on Web has been widely studied and well understood [4, 5, 6]. The main solutions to counter link farming on Web are using iterative ranking algorithms, such as HITS [5] and PageRank [6], to demote spam web pages. Unlike the Web, link farming on Twitter is not among web pages, but social users. Furthermore, each Twitter user could follow others very easily, which causes the cost of building link farming on Twitter is lower than Web's. Therefore, the existing ranking algorithms used on the Web spam could not be directly applied to link farming on Twitter. Although a few algorithms [2, 3] have been proposed to combat link farming on social network, they could not fully take into account relationship strength among users.

In this paper, we firstly introduce a graphical model to estimate relationship strength between users and their neighbors. By utilizing interaction frequency and user similarity, our model reflects the real social phenomenon as well as possible. We then propose a novel ranking algorithm for Twitter, Anti-TrustRank with Relationship Strength (ATRS), which propagate anti-trust score to penalize users for linking to spammers. To model the relationship strength, we not only use interaction frequency to estimate the relationship closeness of users, but utilize entropy to compute the similarities of users. To the best of our best knowledge, no prior study has used the relationship strength to fight link farming on Twitter. We empirically evaluate the proposed method on a real-world Twitter dataset and describe the good performance of the proposed method.

The rest of this paper is organized as follows. Section 2 reviews the related work on link farming. Section 3 proposes a novel ranking algorithm with relationship strength. Section 4 presents the empirical results on a real-world dataset. Finally, we conclude this paper and present the future work.

II. RELATED WORK

A. Link-farming in the Web

Bharat et al. [7] were the first to discover that "mutually reinforcing relationships between hosts" make the ranking algorithm such as HITS to tend to fail. Gyongyi et al. [8] studied that the optimal structures and interconnection of link farms, and provided effective schemes for the research about spam farm later.

Many solutions have been proposed to fight link farming on Web in the past few years, including link-based method and the method based on link structure and webpage content.

By analyzing the link structure, some researchers proposed many algorithms to combat web spam. These algorithms fall into three categories: trust propagation, distrust propagation and the propagation of combining trust and distrust. The first two algorithms propagate either trust scores through links from a set of good seed pages, such as TrustRank[9] and Topical TrustRank [10], or distrust scores through inverse links from a set of bad seed pages to the entire Web [11]. The last algorithms propagate both trust and distrust scores to demote spam pages on Web [12, 13]. Other researchers utilized the link information and content to counter link farming. They often build a classifier to detect nepotistic links or spam by combining the of link-based and content-based features [14, 15].

B. Link-framing in the Social network

To date, there is a few research on link farming in the social network. Ghosh et al. [2] investigated the link farm on Twitter and discovered that most of spammers' links are farmed from not only other spammers but also some normal users. Meanwhile, they proposed a ranking scheme to fight link farming. Yang et al.[3] found that spammers tend to be socially connected to form a small-world network. They revealed three categories of Twitter users that have close friendships with spammers, including social butterflies, social promoters and dummies. To infer more spammers, they design a algorithm by exploiting social relationships and semantic coordinations.

III. THE PROPOSED ALGORITHM ATRS

In the following part, we first introduce the task how to estimate relationship strength, and then use the proposed ATRS algorithm to rank users and detect spammers.

A. Modeling Relationship Strength

1) Definition of Relationship Strength

The definition of relationship strength is as follows: Relationship strength from user $v^{(i)}$ to $v^{(j)}$ denoted as $z^{(ij)}$ is a weight associated with the directed edge $e^{(ij)}$. In most cases, the weight is asymmetric, i.e., $z^{(ij)} \neq z^{(ji)}$, since relationship strength is not only decided by the similarity, also related with the interaction direction and frequency between user $v^{(i)}$ and $v^{(j)}$.

Fig. 1 is the graphical model representation of the relationship strength model. Using this model, relationship strength can be calculated for each pair of users. This method has been proposed for analyzing topical influential user in our former work [16]. Followed by our analyses of relationships between normal users and social spammers, this model applies to fighting link farm on Twitter as well. Let $\mathrm{s}^{(ij)}$ denotes the similarity vector of $v^{(i)}$ and $v^{(j)}$, and $y_r^{(ij)}$ be the current of m different interactions between $v^{(i)}$ and $v^{(j)}$. $z^{(ij)}$ denotes latent relationship strength between $v^{(i)}$ and $v^{(j)}$. $a_r^{(ij)}$ is a set of auxiliary variables $a_r^{(ij)} = [a_{r1}^{(ij)}, a_{r2}^{(ij)}, ..., a_{rl}^{(ij)}]$ for each interaction $r = [1, 2, ..., m]$.

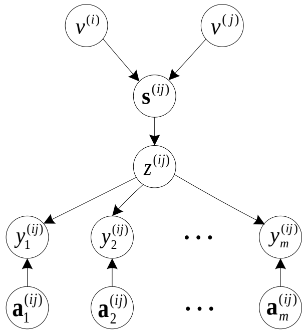

Figure 1. Graphical model representation of the relationship strength model

2) Latent Variable Model

To compute similarity vector among users, user features are introduced firstly. Note that we not only make use of features from our former work, but also extract new features. The summary of these features is listed in Table 1.

TABLE I. USER FEATURES

Category	Feature
Profile	Fofo (following/follower)
	Reputation (follower/following+follower)
	Ntweet (the number of posted tweets)
	Age (the longevity of the account)
Behavior	RURL (URL ratio)
	RU-URL (unique URL ratio)
	Rmention (mention ratio)
	Rhash (hashtag ratio)
	Rretweet (retweet ratio)
	Rdultweet (duplicate tweet ratio)
	Rumention (unique mention ratio)
	Itweet (the interval of posted tweet)
	Ntweet/s (the number of poseted tweets per second)
Neighbor	Anfer (average neighbors' followers)
	Antweet (average neighbors' tweets)
	Anduptweet (average neighbors' duplicate tweets)
	AnURl (average neighbors' URLs)
	AnuURL (average neighbors' unique URLs)
	Anmention (average neighbors' mentions)
	Anretweet (average neighbors' retweetings)

According to homophily theory [17], we assume that the more approximate features' values are, the more similar the pair user is. Hence, we adopt entropy [18] value for all of user features as follows.

For two feature f_i and f_j, we compute similarities with entropy $H(f_i : f_j) = -p\log(p) - (1-p)\log(1-p)$, where $p = \frac{f_i}{f_i + f_j}$.

We proposed a latent variable model, which can be viewed as a hybrid of discriminative and generative model. The joint distribution decomposes as follows:

$$P(z^{(ij)}, y^{(ij)} \mid v^{(i)}, v^{(j)}) = P(z^{(ij)} \mid v^{(i)}, v^{(j)}) \prod_{r=1}^{m} P(y_r^{(ij)} \mid z^{(ij)}) \quad (1)$$

Meanwhile, Poisson regression is employed to solve the generative model. Through the optimized solution of the joint probability, we could obtain relationship strength of each pair of user. (see our optimization algorithm in [16])

B. ATRS Algorithm

We propose ATRS algorithm, an Anti-TrustRank-like approach, to fight link farming on Twitter.The basic idea of our algorithm is to penalize the users who follow social spammers, assuming that a user following spammers is likely to be a suspicious user. More importantly, we consider the relationship strength between social users in the propagation algorithm.

Anti-TrustRank algorithm propagates distrust score via inverse-links from a bad seed set of spammers. In the process of each iteration, the distrust score of a user is averagely partitioned and assigned to the user's followers. This blindfold distrust propagation results in hardness to distinguish spammers from normal users[13]. For example, a spammer who has two neighbors, one is a normal user and the other is a spammer, propagates half of distrust score to both of the neighbors, which causes the two neighbors have the same distrust scores and hard to distinguish them based on their scores. Different from the original Anti-TrustRank algorithm, we propose a more advisable propagation model using relationship strength. In our algorithm, the score *sc(p)* of a user *p* is formalized as follows:

$$sc^{(k)}(p) = (1-\alpha)\cdot d(p) + \alpha \cdot \sum_{q \in out(p)} z^{(pq)} \cdot sc^{(k-1)}(q) \qquad (2)$$

where *d* is the normalized score vector of the bad seed set, *q* is the user who user *p* is following, $z^{(pq)}$ is the relationship strength from *p* to *q* , *out(p)* is the set of links linking out from user *p* and α is a damping factor.

We call the proposed algorithm Anti-TrustRank with Relationship Strength (ATRS), which is described in Alg.1. In ATRS, the score vector d is initialized by setting the set of known spammers to a score, and the rest to 0. The decay factor α is set to 0.85, the most commonly used value.

Algorithm 1 : Anti-TrustRank with Relationship Strength(ATRS)

```
Input: Social Network $G$; Bad seeds set of known spammers $S$;
       Decay factor for ATRS $\alpha$ ; Relationship strength $Z$
Initialize score vector d for notes p in G,
    $d(p) = \begin{cases} \frac{1}{|S|} & if\, p \in S \\ 0 & if\ p \notin S \end{cases}$
$sc^{(0)} \leftarrow d(p)$
Repeat
    Iteratively compute sc according to Formula (2)
Until Convergence
Return sc
Output：Anti-TrustRank with Relationship Strength scores sc
```

IV. EXPERIMENTS

A. Twitter Dataset

In our experiment, we use the real-world Twitter dataset, i.e., UDI Twitter dataset[14]. It contains 140 thousand user profiles, 50 million tweets and 284 million following relationships. Yet, UDI Twitter dataset did not have a ground truth set. So we manually scan the tweets content of all users and click the URLs in the tweets to judge whether they are pornographic information or advertisements. At last, we extracted 1629 spammers and 10450 legitimate users from 12079 users as our dataset, Table 2 shows the statistics of dataset. And then, we extract all features listed in Table 1 based on their profiles and tweets content.

TABLE II. EXPERIMENT DATASET SUMMARY

dateset	**Spammers**	**Normal users**	**Tweets**	**Relationships**
Twitter	1629	10450	1087408	740836

B. Baseline Algorithms

In order to evaluate the effectiveness of ATRS algorithm, we compare with the following baseline algorithms.

1) PageRank[6], which is the basis of most of anti-spam algorithms, is defined as follows:

$$r(p) = (1-\alpha)\cdot \frac{1}{N} + \alpha \cdot \sum_{q \in IN(p)} \frac{r(q)}{|OUT(q)|} \qquad (3)$$

where *q* denotes the user who follow user p and *N* denotes the number of all users in dataset. *IN(p)* is the set of links linking to user *p* and *OUT(q)* is the set of links linking out from user *q*.

2) CollusionRank[2] is a Anti-TrustRank-like algorithm. It is defined as follows:

$$a(p) = (1-\alpha^{'})\cdot d^{'}(p) + \alpha^{'} \cdot \sum_{q \in OUT(p)} \frac{a(q)}{|IN(q)|} \qquad (4)$$

where $d^{'}$ is the distrust score vector of the bad seed set. The difference with our proposed algorithm (ATRS) is that this algorithm averagely assigns the score of *a* user and does not consider the relationship between users.

3) ATRS is our proposed algorithm with relationship strength (see formula (2)).

C. Experimental Results and Discussion

Compared with baseline algorithms, we evaluate our ATRS from the following two aspects: the reputation rankings of spammers and spammer-followers. We selected 80 out of the 1629 spammers as the bad seed set, and computed the ATRS scores for all users in our dataset.

Fig.2 shows that about 55% of the 1629 spammers appear within the top 20% in PageRank, 82% of them are demoted to the last 20% positions in CollusionRank. In ATRS, 85% of all spammers are ranked the last 10% positions. These results illustrate that many sophisticated spammers could achieve higher rankings by constructing link farming with PageRank algorithm, and yet CollusionRank and ATRS could effectively filter out most spammers from the top rankings. Specially, ATRS achieves better performance than CollusionRank.

Fig.3 shows the rankings of the spam-followers using different algorithms. Obviously, the spam-followers, who always are social capitalists, are ranked much higher according to PageRank. About 85% of the spam-followers appear within the top 10% in PageRank. In contrast, CollusionRank and ATRS could effectively demote spam-followers, that is, the users colluded with spammers and other spam-followers would achieve more distrust scores. In Fig.3, we can observe that ATRS outperforms CollusionRank, and it demotes more than 90% of spam-followers to the last 10% positions.

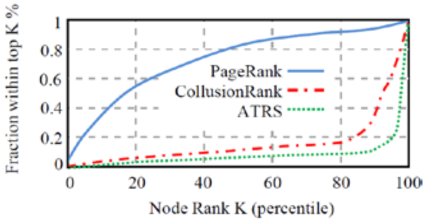

Figure 2. Ranking of spammers

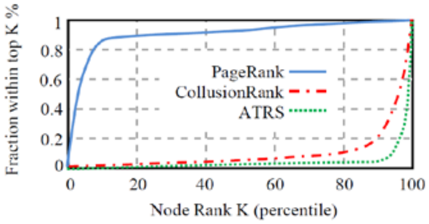

Figure 3. Ranking of spam-followers

These results indicate that not only our algorithm ATRS could successfully fight link farming on Twitter, but also relationship strength could assist ranking algorithm to realize a more advisable scores propagation.

V. Conclusion and Future Works

In this paper, we explore the problem of fighting link farming on Twitter. Our proposed algorithm ATRS takes into account the social relationship strength among users on Twitter. In particular, our proposed Poisson regression-based latent variable model is different from previous methods, which estimated relationship strength by jointly modeling users' similarities and interactions' frequency. The experimental results with a real-world Twitter dataset show that our proposed method is effective and efficient to combat the link farming compared with the state-of-the-art methods.

Next, we plan to extend our work in the following aspects. Firstly, we consider other propagation principles such as TDR[13] for fighting link farming. Secondly, we wish improve our method and realize a combination task of ranking and classification. Lastly, we will attempt to explore the inner structures of the current link farming on Twitter based on the research results.

Acknowledgements

This work was financially supported by National Science Foundation of China (61272374, 61300190), Specialized Research Fund for the Doctoral Program of Higher Education (20120041110046) and Key Project of Chinese Ministry of Education(313011).

Reference

[1] C. Yang, R. Harkreader and G.F. Gu,"Empirical evaluation and new design for fighting evolving Twitter spammers," Journal of Information Forensics and Security, 8(8), p.1280-1293,2013.

[2] S. Ghosh, B. Viswanath and F. Kooti, "Understanding and combating link farming in the twitter social network," Proc. International conference on World Wide Web. ACM, 2012: 61-70.

[3] C.Yang, R. Harkreader and J.L. Zhang, "Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter," Proc. International conference on World Wide Web. ACM, 2012: 71-80.

[4] N.Spirin and J.W. Han, "Survey on web spam detection: principles and algorithms," ACM SIGKDD Explorations Newsletter, 2012, 13(2): 50-64

[5] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM (JACM), 46(5): 604-632,1999.

[6] L.Page, S. Brin and R.Motwani, "The PageRank citation ranking Bringing order to the web," 1999.

[7] K. Bharat and M.R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," Proc. International ACM SIGIR conference on Research and development in information retrieval. ACM, 1998: 104-111.

[8] Z. Gyöngyi and H.Garcia-Molina, "Link spam alliances," Proc. International conference on Very large data bases. VLDB Endowment, 2005: 517-528.

[9] Z. Gyöngyi, H. Garcia-Molina and J. Pedersen, "Combating web spam with trustrank," Proc. International conference on Very large data bases-Volume 30. VLDB Endowment, 2004: 576-587.

[10] B.N. Wu, V. Goel and B.D. Davison, "Topical trustrank: Using topicality to combat web spam," Proc. International conference on World Wide Web. ACM, 2006: 63-72.

[11] V.Krishnan and R. Raj, "Web Spam Detection with Anti-Trust Rank," AIRWeb. 2006, 6: 37-40.

[12] B.N. Wu, V. Goel and B. D. Davison, "Propagating Trust and Distrust to Demote Web Spam," MTW, 2006, 190.

[13] X.C. Zhang, Y. Wang and N. Mou, "Propagating Both Trust and Distrust with Target Differentiation for Combating Web Spam," AAAI, 2011: 1292-1297.

[14] B.D. Davison, "Recognizing nepotistic links on the web," Artificial Intelligence for Web Search, 2000: 23-28.

[15] C. Castillo and D. Donato, "Know your neighbors: Web spam detection using the web topology," Pro. International ACM SIGIR conference on Research and development in information retrieval. ACM, 2007: 423-430.

[16] X.Y. Liu, H. Shen and F.L. Ma, "Topical Influential User Analysis with Relationship Strength Estimation in Twitter," Proc. International Conference on Data Mining Workshop, 2014:1012-1019.

[17] M. McPherson, L. Smith-Lovin and J.M. Cook, "Birds of a feather: Homophily in social networks," Journal of Annual review of sociology, 2001: 415-444.

[18] S. Adali, F. Sisenda and M. Magdon-Ismail, "Actions speak as loud as words: Predicting relationships from social behavior data," Proc. International conference on World Wide Web. ACM, 2012: 689-698.