Detection of Internet Water Army in Social Network

Kun Wang, Yang Xiao, Zhen Xiao
Institute of Network Computing and Information Systems, Peking University, Beijing, China
{wangkun & xiaoyang & xiaozhen}@net.pku.edu.cn

Abstract - As related works had a few research on Internet Water Army in social network, specifically on the Internet Water Army who trends to lead people’s opinions, obscure the real voices and change public opinions in social network. To better understand what difference lie between Internet Water Army and legitimate user, we did some work about behaviour of them from real dataset in Sina microblogging service. We adopted some machine learning algorithms to classify the type of user with collected features through the measurement. At the same time we proposed an influence model and create a new online algorithm with linear complexity to reduce the water army’s influence on legitimate users greatly.

Index Terms - Internet Water Army, Feature measurement, Machine learning, Influence model, MEIWA online algorithm

1. Introduction

As social network websites are popular around the world, more and more people use the social network product like twitter [16], Sina microblogging, facebook and so forth. People spend a lot of time on them to obtain news and information whatever they want. According to the official announcement, the number of Sina microblogging user has exceeded 500 million, daily active users reached 46.2 million by the end of the Feb 2013. With huge number of users and active days, microblogging service attracts a large number of users, but also some others who were specifically hired to publish or disseminate some specific information on social network.

Nowadays these users are called Internet Water Army. The Internet Water Army refers to someone are often hired by PR companies, to post specific content or replies in network and get reward. In social network, the Internet Water Army compared to traditional one is more widely spread and influential.

In our work, we focus on the review posting behavior on microblogging service. We majorly do three aspects of the work. Firstly, the user behavior measurement is done between legitimate user and Internet Water Army. Secondly, the Internet Water Army detection is conducted, and we propose a new online algorithm to reduce the influence of Internet Water Army eventually on microblogging service.

2. Related Work

In some related works, the words like Spammer and Sybil refer to the users who have malicious attacks, distorted the true, disseminating irrelevant advertising and so forth [2].

Since spammers affect the normal use of the Internet users badly. For example, they may make some fake review and fake scoring[8], overstate the effect of product, post some malicious comments against the competitors [5] or duplicate comments in the forums, publishing articles and irrelevant reviews, disseminate irrelevant advertising [2]. In social networking sites, they also spread malicious links, publish irrelevant ads and content [10].

Generally people take some methods to deal with related problems like machine learning [15] algorithms for detecting spammers [9,10], detecting the publishing time distribution to find out abnormal situations [4,12], take advantage of the relationships between users in social network data for pattern discovery or community detection [5].

In previous work, spammer detections are mostly focused on the publication of malicious URLs [13], irrelevant advertising user detection or measuring user behaviors [10].

To the best of our knowledge few work is done on Internet Water Army detection in social network. We are also the first to propose a method to measure the influence of the Internet Water Army and a new algorithm to reduce the influence.

3. Internet Water Army Detection

In Sina microblogging service, PR companies often hire Internet Water Army to retweet or reply tweet to support or against some specific incident [8], attack competitors, cover up the truth and obtain public’s supports and sympathy. In psychology theory, people have a conformity mentality [1], the personal behavior in the crowd affected by the outside world, people prefer to the same performance with majority of people around them in their perception, judgment and cognitive. Some PR companies take advantage of the people's psychological law to hire Internet Water Army to create majority of a fake people's point in outward seeming, to stop legitimate users expressing their real voice and affect their point of specific event. The contributions of our work can be summarized by as followings:

Firstly, we measure the both legitimate users and Internet Water Army‘s behavior from different perspectives. We use a variety of machine learning-based classification algorithms to detect Internet Water Army with a precision more than ninety-two percent. Finally we proposed a new model to measure how Internet Water Army affects the legitimate users. We reduced Internet Water Army’s influence to legitimate users one-sixth in average with a new online algorithm

A. Dataset

We crawled some popular tweets whose comment number is over 7000. We picked up 307 tweets from them with all comments and reviewers’ personal information and relationships in Sina microblogging service. We collected 4,000,000 comments, more than 100 million unique users and personal relationships in total. Then we manually labeled the microblogging users to legitimate user or Internet Water

Army. We labeled the data by looking up the personal information, home page, photos albums, comments content to determine whether the user is Internet Water Army or not. Judgment is based on whether the publication is exaggerated, self-contradicted [7], personal user information is fake or user rating is low. A total number of 212 Internet Water Army and 732 legitimate users were labeled.

B. Feature Measurement

Based on manually labeled data, we measured users' behaviors through the ratio of users' bilateral friends, the ratio of users' retweet and users' microblogging ranks to see what difference lie between legitimate users and Internet Water Army.

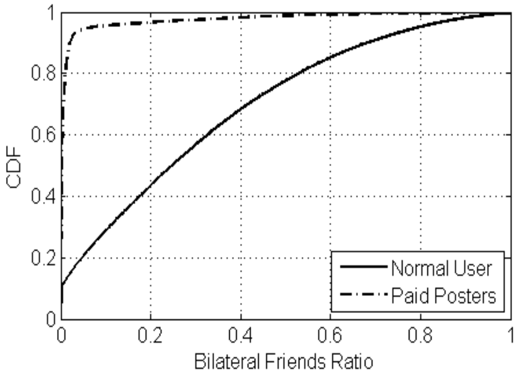

Fig. 1 Bilateral Friends Ratio CDF

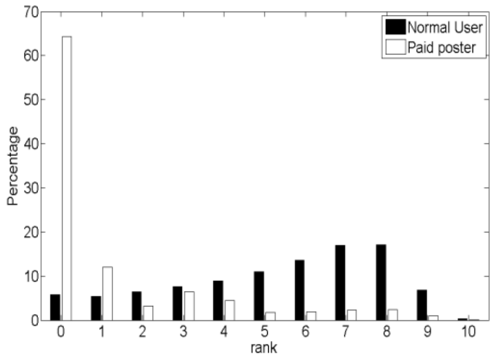

Fig. 2 User Microblogging rank distribution

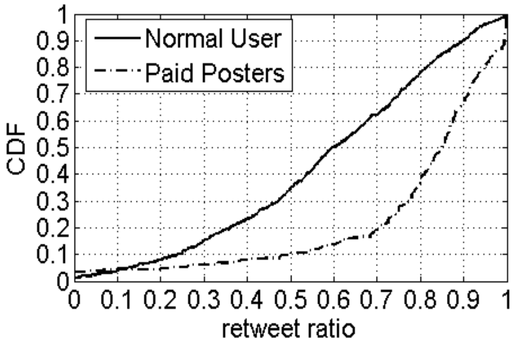

Fig. 3 Retweet Ratio CDF

From Fig 1 we can see Internet Water Army's bilateral friend ratio is significantly lower than legitimate users'. The reason is that the feature of bilateral friends is usually based on relation in real world. People already know each other offline are more likely add friends with each other in social networks. But Internet Water Army doesn't care about the real relationship in social network. This is the main reason why Internet Water Army's bilateral friend ratio is much lower.

In Fig 2, Internet Water Army's microblogging user rank is lower than legitimate user. Because the Internet Water Army usually doesn't care about what they had posted, the account online time and so on to get higher user rank.

In Fig 3, the Internet Water Army's retweet ratio is much lower than legitimate users'. The legitimate users prefer to post tweets by themself to present their opinions or feelings instead of focusing retweeting to spread others' opinions on purpose.

C. Classification

By looking up the data in the microblogging service, some comments between the users or the user self are repeated or quite similar. In the paper [7] published in 2008, the author also mentioned some conclusions, which indicated these similar comments posted by different users are very suspicious. So we define a comment similarity feature.

Firstly, we defined the similarity between different comments with the Jaccard coefficient. We calculated similarity between all comments with length over five Chinese characters [14].

$$\mathrm{Jaccard}(c1, c2) = \frac{S(c1) \cap S(c2)}{S(c1) \cup S(c2)}$$

$$V(c) = \mathrm{Max}\big(\mathrm{Jaccard}(c, c1)\big) \ \wedge c1 \in A - c1$$

$$\mathrm{Sim}(u) = \mathrm{Max}(V(c)) \wedge c \in C(u)$$

As shown in the above formulas, S(c) is collection of ratings c which was generated by the method of bigram segmentation. And then we used the Hash function for fast matching [12] to calculate both set intersection and union. Jaccard (c1, c2) indicates the Jaccard coefficient between comments c1 and c2. V (c) represents the maximum similarity among others comments. Sim (u) represents the user u's maximum similarity among his all comments' similarity M(c).

Secondly, we calculate the ratio of retweets over all tweets. The RetweetRio(u) represents whether a user prefer original tweet over retweeting

Thirdly, according to the experimental results, when a lot of comments are published in a short period of time, it means Internet Water Army is doing their works in social network site [3], posting a massive number of comments. So we defined ComRio(u) which indicates the average of the number of comments posted by user u during the Δt time.

$$\mathrm{comRio}(u) = \frac{\mathrm{Avg}\big(\mathrm{Count}(c)\big)}{\Delta t} \ c \in \mathrm{Comment}(u)$$

$$\mathrm{Count}(c) = \#\mathrm{Comment\ at}\ [ct - \Delta t, ct]$$

Fourthly, we collected user rank from user information from microblogging service as Rank (u). Microblogging user rank is officially defined which indicates the user online time and daily continuous login time. In order to get a higher rank, user must login microblogging account for a long time or post original tweet. Obviously legitimate users have higher ranks than Internet Water Army from the figure 2.

Fifthly, we collected users' bilateral relationship as Bilateral(u) which indicates if the user builds regular relationship in social network sites as people often build their social networks from real world relationships.

After above features were normalized, we adopted a data analysis tools named Weka [6] for a classification experiment using cross-validation method to verify the classification results' precision and recall.

TABLE 1. Classification Results

Classfication Algo	Precision	Recall	F Value
NaiveBayes	0.898	0.894	0.895
J48	0.919	0.919	0.917
Logistic	0.926	0.926	0.926
VotedPerceptron	0.928	0.929	0.928

As shown in Table 1, when we adopted all features for classification, the precision and F value are both over ninety percent in average with four kinds of classification algorithms. Which means the features what we had chosen are appropriate for the classification problem.

To prove the different features bring the different results in the classification, we made different combinations of the features. In order to control variables, all combinations used Logistic algorithm for classification experiments.

TABLE 2. Feature Combination of Classification Results

Features combination	Precision	Recall	F Value
[Rank,Bilateral,Sim}	0.916	0.917	0.917
[Rank,Bilateral}	0.89	0.892	0.891
{Sim,ComRio,RetweetRio}	0.843	0.848	0.844
{Sim,ComRio}	0.808	0.813	0.81
{ComRio,RetweetRio}	0.737	0.762	0.728

As shown in Table 2, with only Rank, Bilateral, Sim three features in the classification algorithm can achieve 0.91 accuracy and recall. We will use this combination for the new linear complexity algorithm's features.

4. Online Algorithm

A. Influential Model

The experiment results showed the classification algorithm worked well. But we can only detect one tweet if it is attacked by Internet Water Army after the attack. During this time, the legitimate users had been affected by the Internet Water Army.

Therefore, we proposed a new method to measure the how much Internet Water Army had impacted the legitimate users and a new online algorithm to reduce the impact. As the comments are ordered by time, so people always see the latest comments about one tweet. Based on the situation, we proposed a model to evaluate the Internet Water Army's impact on legitimate users.

$$\mathrm{tc} = \#\mathrm{Comment\ at}\ [\mathrm{ct} - \Delta\mathrm{t}, \mathrm{ct}]$$

Suppose the user u post a comment at time ct, then we can get the total number of comments tc during the fixed time window Δt during [ct-Δt, t]. Among these comments, sc indicates the number of comments Internet Water Army had posted. So we derived the impact factor p like:

$$\mathrm{p} = \frac{\mathrm{sc}}{\mathrm{tc}}$$

By calculating the impact factor p we can measure how many comments posted by legitimate users. As the factor defined, the more Sybil comments user have viewed, the factor is larger.

B. Online Algorithm

By feature extraction and machine learning could achieve a high precision classification to distinguish Internet Water Army and legitimate users. However, the text similarity calculation is often essential to wait until the end of the event calculated off-line, and the high complexity is unbearable.

```
Minimizing the effect of Internet water Army
$\mathrm{i} \leftarrow 0$
while $|\mathrm{Q}| < WindowSize\ do$
$\mathrm{Q.push\_back(C(i))}$
$\mathrm{i} \leftarrow \mathrm{i} + 1$
end while
while $\mathrm{i} < |\mathrm{C}|$ do
$\mathrm{AvgRank} \leftarrow \sum_{\mathrm{c}\in\mathrm{Q}} \frac{\mathrm{Rank(c)}}{|\mathrm{Q}|}$
$\mathrm{pRank} \leftarrow \frac{\mathrm{Rank(C(i))}}{\mathrm{Max(Rank\big(C(i)\big), AvgRank)}}$
$\mathrm{AvgBiRatio} \leftarrow \sum_{\mathrm{c}\epsilon\mathrm{Q}} \frac{\mathrm{BiRation(c)}}{|\mathrm{Q}|}$
$\mathrm{pBiRatio} \leftarrow \frac{\mathrm{BiRatio(C(i))}}{\mathrm{Max(BiRatio\big(C(i)\big), AvgBiRatio)}}$
$\mathrm{p} \leftarrow \sqrt{\mathrm{pRank} \cdot \mathrm{pBiRatio}}$
if random(p) then
        Queue.pop_front()
        Queue.push_back(C(i))
end if
  $\mathrm{i} \leftarrow \mathrm{i} + 1$
end while
```

Algorithm 1. MEIWA online algorithm

According to Table 2, we can see that the Rank and Bilateral group has a high accuracy relatively, and these two features doesn't need to be calculated in advance in which is easy to get in a constant time. So we proposed an algorithm using these two features to measure a person's credibility comment. The Algorithm 1 ensures the higher reliability comments more likely to remain in the user’s visible time window. At the beginning of the algorithm a queue is used to save the trusted comments in current time window. Whenever a new comment is generated, it will calculate the ratio of user's rank and the average user rank in the queue, pRank as well as the ratio of user’s bilateral friends and the average of the queue’s pBiRaito. Then we calculate the value p as the square root of both parameters. The value p is the possibility of the comments whether it should be pushed in the comments queue to be viewed by uses. Algorithm 1 is an online algorithm, which ensure the order of time and effectively reduced the influence of Internet Water Army.

TABLE 3. MEIWA Online Algorithm Results

	Sequence strategy	**MEIWA Algo**
average	0.738	0.122
variance	0.758	0.14

From Table 3, Under the MEIWA algorithm, the influence coefficient of Internet Water Army reduced from 0.738 to 0.122, reduce the effect to one sixth. Meanwhile, the algorithm in linear time guarantees the reviews ordered by time which is consistent with the users’ reading habits. As the new MEIWA algorithm treats the quality of tweet by each user as the main reason to keep the tweet in the watching queue. So uses are more likely to see tweet posted by legitimate uses as queue’s property first in first out keep the sequence of posting.

5. Conclusion

In this paper we measured the Internet Water Army’s behavior from multiple dimensions. Then we selected several effective features as the training model and use machine learning methods for classification. Based on the behavior of users viewing the comment, we proposed a model to measure the influence of Internet Water Army. In order to reduce the influence coefficient, we proposed a new linear time complexity online algorithms named MEIWA. The new algorithm results showed that the influence is reduced to one sixth of the sequence strategy which is used by default with ensuring users’ viewing comments habits. In this paper, there are still a lot of future works to do:

(1) The Internet Water Army also likes to forward the tweets to expand their influence on microblogging service. With forwarding behaviors are more widely viewed by people, we can proceed with a more comprehensive understanding of the Internet Water Army’s behavior.

(2) With the deep into Internet Water Army’s more detailed behaviors, we found that there are much more classifications in themselves. For example, some water army is only responsible for supporting some super stars. And some of them were devoted to the social phenomena event to post some tweets or give a review. For the different kinds of Internet water Army, we would like to do some research about how much influence they really have.

References

[1] Asch, S.: Studies of independence and conformity: I. a minority of one against a unanimous majority. Psychological Monographs: General and Applied (1956)

[2] Benevenuto F, Magno G, Rodrigues T, et al. Detecting spammers on twitter//Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS). 2010.

[3] Chen C, Wu K, Srinivasan V, et al. Battling the internet water army: Detection of hidden paid posters. arXiv preprint arXiv:1111.4297, 2011.

[4] Feng S, Xing L, Gogar A, et al. Distributional footprints of deceptive product reviews//Proceedings of the 2012 International AAAI Conference on WebBlogs and Social Media, June. 2012.

[5] Gao H, Hu J, Wilson C, et al. Detecting and characterizing social spam campaigns//Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. ACM, 2010: 35-47.

[6] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10-18.

[7] Jindal N, Liu B.Opinion spam and analysis//Proce-edings of the international conference on Web search and web data mining. ACM, 2008: 219-230.

[8] Lim E P, Nguyen V A, Jindal N, et al. Detecting product review spammers using rating behaviors//Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010: 939-948.

[9] Ott M, Choi Y, Cardie C, et al. Finding deceptive opinion spam by any stretch of the imagination. arXiv preprintarXiv:1107.4557, 2011.

[10] Thomas K, Grier C, Song D, et al. Suspended accounts in retrospect: an analysis of twitter spam//Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, 2011: 243-258.

[11] Wang G, Wilson C, Zhao X, et al. Serf and turf: Crowdturfing for fun and profit//Proceedings of the 21st international conference on World Wide Web. ACM, 2012: 679-688.

[12] Yang C, Harkreader R, Zhang J, et al. Analyzing Spammers’ Social Networks for Fun and Profit. Proc. World Wide Web, 2012: 16-20.

[13] C. Aggarwal. On abnormality detection in spuriously populated data streams. In Proceedings of the 5th SIAM Data Min. Conference, 2005.

[14] G. W. Alpers, A. J.Winzelberg, C. Classen, H. Roberts, P. Dev, C. Koopman, and C. Barr Taylor. Evalua-tion of computerized text analysis in an internet breast cancer support group. Computers in Human Behavior, 21(2):361{376, 2005.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993{1022, 2003.

[16] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In Proceedings of the 19th international conference on World wide web. ACM, 2010.