EasyChair Preprint

№ 735

Detection Method for User Click Fraud Based on Garbled Bloom Filter

Chunliang Zhou and Le Wang

January 18, 2019

Detection Method for User Click Fraud Based on Garbled Bloom Filter

Chunliang Zhou[1] and Le Wang[1*]
[1]Ningbo University of Finance & Economics, Ningbo,China
chunliangzhou@nbdhyu.edu.cn,wangleboro@163.com

Abstract

In order to solve the click fraud problem in the network, a detection method of user click fraud is presented with confusion Bloom filter. In this method, the user click fraud advertising set is found by the training data set attributes at first, and then use the Bloom filter and outlier mining algorithm for the detection and localization of suspected fraud. Finally, all suspected fraud members were classified by the Bias classification method to detect fraud. The experiment results show that this method can effectively shield visitors from accidental unconscious clicks, and significantly reduce the probability of click fraud.

Keywords: Click Fraud; Bloom Filter; Bias; Outlier Mining

1 Introduction

In recent years, as the Internet technology is continuously life-oriented, and the online advertising is developing rapidly, and the cost per click mode of online advertising is emerging gradually, “fraudulent click” behavior has become a major problem that puzzles advertisers and service providers[1-5] Fraudulent click refers to all click behaviors by fraudulent means or with fraudulent intent and acknowledged by search engine, which greatly harms the healthy development of Internet advertising, so the detection and identification of fraudulent click behavior plays a very important role

in guaranteeing the security of user information.

At present, main prevention and detection methods adopted at home and abroad are as follows: (1) anti-fraud click method based on IP and Referer detection means; (2) three filtration technologies, i.e. click analysis filtration, historical behavior analysis filtration and AI mode identification filtration; (3) real-time detection method based on finite state automaton (DFA); (4) anti-fraud click method based on graphic verification code, etc. Currently, the feature selection method for user behaviors is mainly adopted for the detection of click fraud users, and characteristic dimensionality reduction methods that are frequently used are as follows: primary component analysis (PCA), multiple dimensional scaling (MDS), linear discriminant analysis (LDA), etc. The feature selection method for fraud detection presented in reference [6] verifies that the fraud detection system subject to feature selection can effectively detect the click fraud users. Reference [7] proposes a way to count the feedback data by the use of inserted false advertisements to help judge the visitors as a third party. Reference [8] proposes a method to identify the online water army groups in e-commerce field, which can analyze the candidate groups through comment contents and discover the online water army groups according to their behavior characteristics. In addition, fraudulent clickers appear in groups, and this group click behavior is easy to discover[9-11]. Reference [12] proposes that the classification algorithm of support vector machine (SVM) is used for fraud detection system, which can detect the fraudulent behaviors in a better way. Reference [13] proposes to use LibSVM as SVM training and testing tool, and mine the suspected fraudulent groups with Apriori algorithm, and analyze the attributes of user click behaviors in a group, and find out the suspected fraudulent users who have significant difference from other users in the group using outlier mining method, and classify and analyze the suspected fraudulent members with Bayesian classification method to get the real fraudulent users. However, the detection method for user click fraud proposed at present has higher requirements for the hardware and network of participants, and the resource cost is relatively large, and the detection rate and accuracy are difficult to meet the real-time requirements.

On the basis of the work above, a new detection method for user click fraud is presented in this paper based on garbled bloom filter, and the performance difference between this method and other methods is compared and studied by combining simulation experiments. The structure of this paper is as follows: Section 1 describes the evaluation index of user click fraud detection method; detection method is established in Section 2 using garbled bloom filter; simulation experiments are carried out in Section 3; Section 4 summarizes the full text.

2 Evaluation Index

Each influence factor for the preliminary evaluation of user click stream has its own weighted

value ω_i（$0 \le \omega_i \le 1$） and attribute value r_i（$0 \le r_i \le 1$）, and is weighted to a total evaluation score *S*:

$$S = \frac{\sum_{i=1}^{n} \omega_i r_i}{\sum_{i=1}^{n} \omega_i r_i + \sum_{i=1}^{n} (1 - \omega_i r_i)} \quad (1)$$

The k-distance of any object *p* in the click stream data set is the maximum distance from *p* to its nearest neighbor, which is written as k-distance(*p*). The *k*-distance neighborhood of the object *p* is written as $N_{k\text{-}distance}(p)$, which contains all objects whose distance is not greater than *k*-distance of *p*. The reachable distance from the object *P* to the object *o* (where o is in the *k* nearest neighborhood of *p*) is defined as:

$$read\_disk_k(p,o) = \max\{k - dis\tan ce(o), \|p - o\|\} \quad (2)$$

Local reachability density ($lrd_k(p)$) of *p* is defined as the reciprocal of average reachability density of *k* nearest neighbor points based on *p*:

$$lrd_k(p) = \frac{\left|N_{k\text{-distance(p)}}(p)\right|}{\sum_{o \in N_{k\text{-distance(p)}}} reach\_disk_k(p,o)} \quad (3)$$

The representation *p* of local outlier factor *LOF* of *P* is set as the degree of outlier:

$$LOF_k(p) = \frac{\sum_{o \in k-dis\tan ce_{(p)}(p)} \frac{lrd_k(o)}{lrd_k(p)}}{\left|N_{k-dis\tan ce(p)}\right|} \quad (4)$$

$I=\{i_1,i_2,\ldots,i_n\}$ is the set of *n* different items. If $X \subseteq I$ and $k=\left|X\right|$ in a set X, then *X* is the *k* item set, or an item set. *D* is the set of affair *T*, $T \subseteq I$. For the given database *D*, the support degree of *X* is defined as the number of *X* affairs included in *D*, which is written as sup(X). An user can customize a minimum support degree less than $\left|D\right|$, which is written as min_s. The affair database *D* and support degree min_s are given. For the item set $X \subseteq I$, if sup(X) $\ge$ min_s, then *X* is the frequent item set in *D*; if sup(X) $\ge$ min_s and sup(Y)<min_s for $\forall(Y \subseteq I \wedge X \subset Y)$, then *X* is the maximum frequent item set in *D*.

Click stream data sets have many attributes and tuples, and each data set corresponds to a hash function. The legality of user's click behavior is predicated with Bayesian classification method, and

the legality of user click is set as T. T_i is the scope of click behavior trust, and $|T_i|(1 \le i \le L)$ is the number of times that the overall trust in the click history of the predicated click user is in the scope of T_i. X is a click event, namely:

$$X = \{x_1, x_2, ...x_{\max}\} \tag{5}$$

Where *max* is the maximum number of attributes of a click event. The prior probability of legal level of user click behavior is:

$$p(T_i) = \frac{|T_i|}{n}(1 \le i \le L)\left(\sum_{i=1}^{L} p(T_i) = 1\right) \tag{6}$$

Where n is the total number of previous clicks of the predicated click user. Assuming that the values of each attribute are independent, and the prior probability is $p(X_1|T_i)$, then $p(X_2|T_i)...p(X_n|T_i)$ can be obtained from the training data set, and the prediction probability of legal level of a click behavior of the user is $p(X|T_i)p(T_i)$.

A bloom filter is the array of a m bit, which can represent a set S with ω elements at most. k hash functions $H = \{h_1, h_2, ..., h_k\}$ that are independent of each other and uniformly selected are used, and the set S has n elements, $S = \{s_1, s_2, ..., s_n\}$, which are mapped to k corresponding values through k hash functions. Assuming that the hash function is uniformly distributed, after all the elements in the set are mapped, the probability that any bit of the bloom filter is zero is:

$$p = \left(1 - \frac{1}{m}\right)^{kn} \approx e^{-kn/m} \tag{7}$$

When the elements that do not belong to the set are misjudged as belonging to the set, the value that needs to meet each corresponding bit is 1, that is, the misjudgment rate of the element is:

$$f^{BF}(m,k,n) \approx (1-p)^k \tag{8}$$

Namely:

$$f^{BF}(m,k,n) \approx (1-p)^k = \left(1 - e^{-kn/m}\right)^k = \exp\left(k \ln\left(1 - e^{-kn/m}\right)\right) \tag{9}$$

If the upper limit f_0 of misjudgment rate is specified, the maximum number of elements represented by the filter can be calculated from the above equation when the filter length *m* and hash function *k* are fixed:

$$n_0 = -\frac{\ln\left(1-e^{\ln f_0/k}\right)\cdot m}{k} \tag{10}$$

As can be seen from $g(k) = k\ln\left(1-e^{-kn/m}\right)$, the functions *g* and *f* can reach the minimum value at the same time, and the derivative of *k* is taken for *g* to get:

$$\frac{dg(k)}{dk} = \ln\left(1-e^{-kn/m}\right) + \frac{kn}{m}\frac{e^{-kn/m}}{1-e^{-kn/m}} \tag{11}$$

If $\frac{dg(k)}{dk} = 0$, then:

$$k_{\min} = (\ln 2)\left(\frac{m}{n}\right) \tag{12}$$

When *k* satisfies the equation (12), the minimum misjudgment rate is obtained.

As the user click has many redundant characteristics in practical application, there is a certain error in parameter estimation only relying on the above methods. Therefore, this paper proposes a detection method for user click fraud based on garbled bloom filter to correct the above calculation results and thus improve the detection accuracy and efficiency.

3 Detection Method Based on Garbled Bloom Filter

Bloom filter (BF) is a spatially efficient probability data structure representing bit strings[14, 15], which supports the hash query of elements and can be used to test whether an element x is included in the set *S*, and can meet the efficient storage and query requirements of resources. The essence of its algorithm structure is to map all elements in the set to the bit string vectors through *k* hash functions. Different from the traditional hash storage table, the hash table is degenerated into a bit string vector *V* in the bloom filter, and each element only occupies a few bits. Standard bloom filter has such problems as high misjudgment rate and inaccurate data processing in the process of data query, so the sets cannot be directly mapped to the bloom filter for storage and query. To solve the above problems,

the bloom filter is improved using the outlier mining algorithm[16] to establish the garbled bloom filter for data storage and comparison.

The dissimilarity of user attribute is defined here: when the attribute is discrete data, the distance metric is 0 only when they are exactly equal, otherwise it is 1. The dissimilarities of two click records x，y are defined:

$$D(x,y)=\sum_{f\in Fields}\frac{\omega_f}{d_f(x,y)} \tag{13}$$

The click log records of each user are composed of vector forms of click attributes (user IP, user source URL, region, query term); *Fields* is the feature set in click attribute; $d_f(x,y)$ is the distance measure of each f (attribute), i.e. [0,1]; w_f is the weight of $d_f(x,y)$. For the user source URL attribute, the distance measure can be defined as:

$$d_{URL}(x,y)=1-\frac{LCP(x,y)}{\max(|x|,|y|)} \tag{14}$$

Where *LCP* is the longest common prefix of two click log records. Assuming that the i^{th} continuous attribute of two data sets x and y on heterogeneous data set X is x_i and y_i respectively, then the distances of x and y on the i^{th} attribute are:

$$\text{normalized diff}_i(x,y)=\frac{|x_i-y_i|}{4\sigma_i} \tag{15}$$

Where σ_i is the variance of the i^{th} attribute on the data set.

Meanwhile, the value difference metrics of x and y on the j^{th} attribute are:

$$\text{normalized } vdm_j(x,y)=\sqrt{\sum_{j=1}^{C}\left|\frac{N_{j,x,c}}{N_{j,x}}-\frac{N_{j,y,c}}{N_{j,y}}\right|^2} \tag{16}$$

Where $N_{j,x}$ is the number of data that the value of the j^{th} attribute of all data on the data set X is x_i, and $N_{j,x,c}$ is the number of data that the value of the j^{th} attribute of all data on the data set X is x_j and the output category is C, and C is the data output category.

In addition, the distance function $H(x,y)$ of heterogeneous value difference metric (HVDM) defined here is:

$$H(x,y)=\sqrt{\sum_{i=1}^{m} d_i^2(x_i,y_i)} \tag{17}$$

When x_i or y_i value is null, the value is 1; when x_i or y_i belongs to continuous attribute, the value is calculated as per the equation (15); when x_i or y_i belongs to discrete attribute, the value is calculated as per the equation (16).

According to the above description, the detection algorithm for user click fraud based on garbled bloom filter is presented here using outlier mining method:

Step1: the parameters are initialized, and the data are pre-processed;

Step2: the length of bloom filter is set as m, and k hash functions $H=\{h_1,h_2,...,h_k\}$ that are independent of each other and uniformly selected are used, and the threshold of misjudgment rate is λ;

Step3: the set of attribute features of user click fraud advertising is found through training data set to get the evaluation score s from the equation (1);

Step4: the minimum probability of legal level is set as n, and the legality of click behavior is detected as per the equations (6) and (7). If it is less than n, please return to Step1, or jump to Step 5;

Step5: the outlier mining based on the dissimilarity is carried out:

1. The dissimilarity matrix $D(x,y)$ is initialized;
2. The click user set in the click stream data set S_f is scanned;
3. The dissimilarity between users is calculated based on the dispersion of distance metric, and the dissimilarity matrix $D(x,y)$ under all attribute conditions is obtained as per the equations (13) and (14);
4. A certain number k of outliers, i.e. suspected fraudulent users in S_f are filtered to get the candidate fraud group S_c.

Step6: the candidate fraud group S_c obtained in Step5 is calculated as per the equations (8) and (9) to get the misjudgment rate; if the misjudgment rate is less than the threshold λ, please return to Step 7, or jump to Step 3;

Step7: the algorithm ends.

4 Simulation Experiment

The simulation analysis is conducted based on MATLAB in this paper to verify the effectiveness of the detection method above. This paper considers the weight of each attribute in the click stream data set as the same level, and the attributes include (user IP, user source URL, region, query term).

Meanwhile, the data set is selected: 100 users are randomly selected for experiments, and 40 students are randomly selected from them as fraudulent users. The experiment period is 2 weeks, and a total of 50000 click log data are collected. Pure Bayesian classification method, Google detection method and detection method proposed in this paper are first compared in this experiment. Table 1 shows the number of fraudulent clicks of three detection results.

Total number of click log records	Detection results of Bayesian classification method	Google detection results	Detection results of the method proposed in this paper
6409	1689	1873	1786
6306	1511	1783	1598
7289	2987	3266	3321
7403	1727	1909	1824
6981	993	864	787
7995	2587	2457	2402
8217	3375	3112	3216

Table 1: Comparison of Detection Results

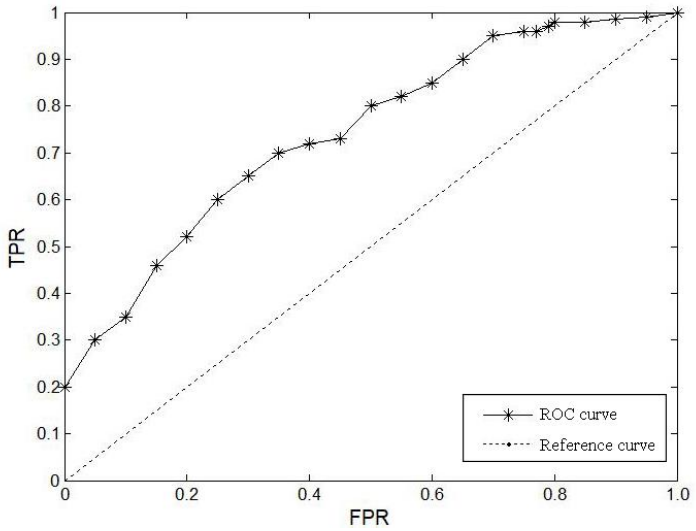

Fig. 1 Comparison of ROC Curve and Reference Line

As can be seen from Table 1, the detection results of the method described in this paper are relatively close to that of Google. Since the data provided by Google is of high accuracy, the reliability of the method described in this paper is also verified. Meanwhile, ROC curve is used in this paper to evaluate the accuracy of the method proposed in this paper. ROC curve is a curve mapped by taking true positive rate as the vertical coordinate and false positive rate as the horizontal ordinate, in

which true positive rate TPR=TP/TP+FP, and false positive rate FPR=FP/TN+FP. TP is the number of positives judged as positives, and FP is the number of negatives judged as positives. Fig. 1 shows the ROC curve of detection method proposed in this paper, and the farther the curve is from the reference diagonal, the better the detection effect. As can be seen from Fig. 1, the value of the area under the curve (AUC) is close to 0.9, which indicates that the method proposed in this paper has a higher accuracy in detecting the user fraud. When AUC is greater than 0.5, the closer the AUC is to 1, the better the detection effect. If AUC is equal to 0.5, the detection method is completely ineffective, which indicates that the detection method is useless.

In the algorithm proposed in this paper, the support degree sup (support degree=*u**number of log records, *u* value is 0~100%) will inevitably affect the number of suspected fraudulent users, and thus affect the accuracy of detection results. Fig. 2 and Fig. 3 respectively show the impact of the variation of *u* value on the number of suspected fraudulent users and accuracy rate. As can be seen from Fig. 2, with the increase of *u*, the number of suspected fraudulent users presents a monotone decreasing change. When *u* is equal to 0, the number of suspected fraudulent users reaches a peak, and all fraudulent users are detected; when *u* is equal to 100, the number of users in the suspected fraud group is 0. This is because that no user can click all the advertisements within a certain time. As can be seen from Fig. 3, when *u* value is 40%~50%, the support degree value is the optimal value, and the accuracy rate will reach a peak, approaching 90%. With the increase of *u*, the accuracy rate decreases and the rate gradually slows down, which shows a long-tail distribution. When *u* takes the maximum value, namely the support degree is the maximum, the accuracy rate is close to 0.

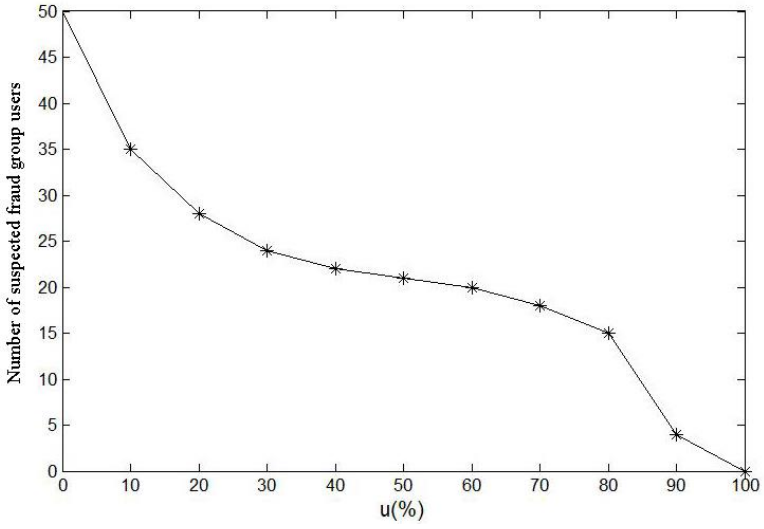

Fig. 2 Variation Relationship between Support Degree and Number of Suspected Fraud Group Users

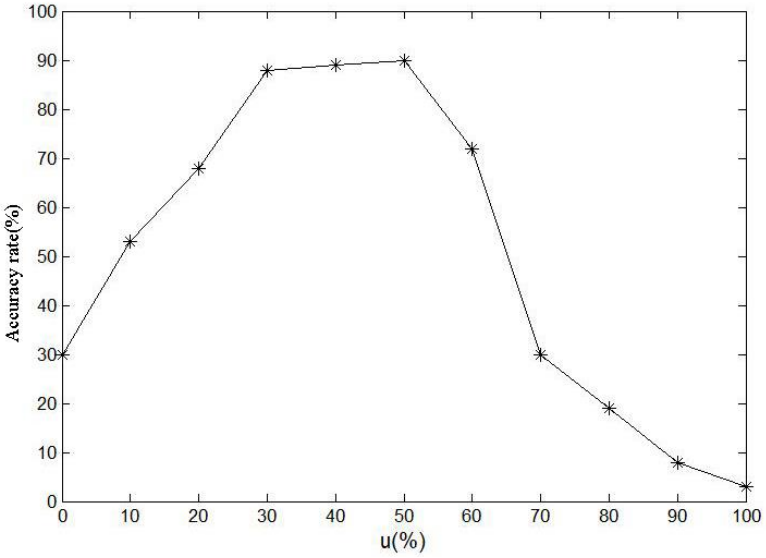

Fig. 3 Variation Relationship between Support Degree and Detection Accuracy Rate

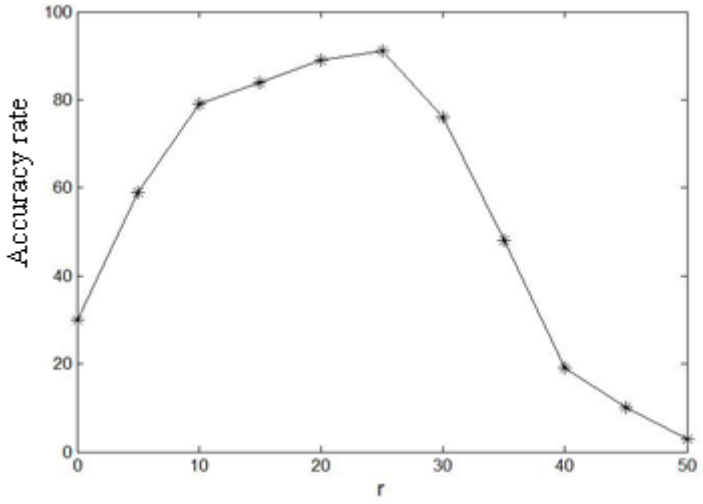

Fig. 4 Variation Relationship between the Parameter r and Detection Accuracy Rate

In addition, Fig. 4 shows the variation relationship between the parameter *r* and accuracy rate of detection method under the condition that the value of support degree sup remains unchanged. As can be seen from Fig. 4, when the *r* value is 25~30, the accuracy rate will reach a peak, approaching 90%. This is because the experiment assumes that the number of non-fraudulent users is 30, so when the *r* value is around 30, non-fraudulent users will be basically filtered out and the detection accuracy will

be the highest. However, when the *r* value is close to 50, the detection accuracy is close to 0 as the users in the suspected fraud group (including fraudulent users) are filtered out.

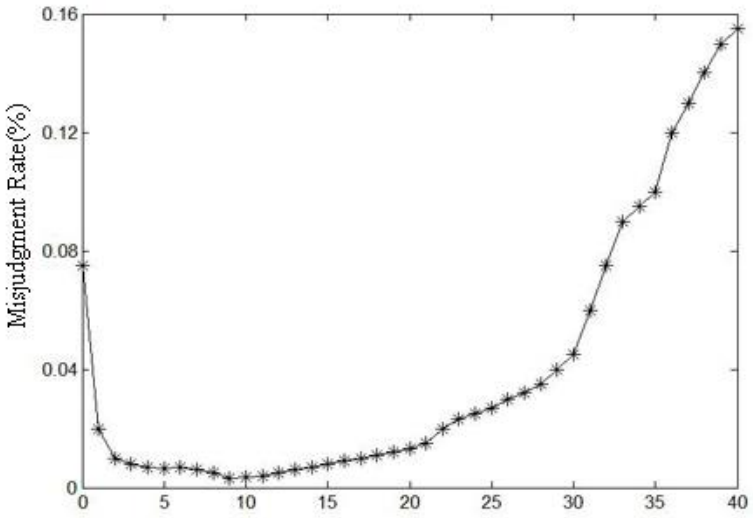

Fig. 5 Variation Relationship between the Number *k* of Hash Functions and Misjudgment Rate

The number *k* of hash functions also has an impact on the misjudgment rate of the detection method. The filter length is set as 2560bit, and the number *k* of hash functions k is set as (0~10). Fig. 5 shows the variation curve of the detection method proposed in this paper over the number *k* of hash functions. As can be seen from Fig. 5, when the number *k* of hash functions is 10, the misjudgment rate of the detection method is the lowest. When *k* exceeds the specified range (0~10), the misjudgment rate is positively correlated with the number *k* of hash functions. When the number *k* of hash functions is 0~10, the more the number is, the smaller the misjudgment rate of the detection method. This is because in the specified range (0~10), the more the number of hash functions is, the more bits mapped by elements in the vector and the more the element information expressed, so the misjudgment rate is reduced.

5 Conclusion

As the detection method for user click fraud proposed at present has higher requirements for the hardware and network of participants, and the resource cost is relatively large, and the detection rate and accuracy are difficult to meet the real-time requirements, a detection method for user click fraud based on garbled bloom filter is proposed based on outlier mining method to achieve efficient detection. Finally, the key parameters that affect the detection efficiency, accuracy rate and

misjudgment rate are compared and analyzed through simulation experiments, and the detection method is improved through parameter setting, which thus significantly reduces the probability of user click fraud. In the follow-up study, other artificial intelligence algorithms and evaluation indexes can be considered to improve the detection method for user click fraud.

References

[1] Hao Y, Bin W, Gang X, Xiaochun Y.(2010). Distance-Based Outlier Detection on Uncertain Data. *Journal of Computer Research and Development*(pp.474-484).

[2] Qiao L, Hui H, BINxing F.(2014). Awareness of Network Group Anomalous Behaviors Based on Network Trust. *Chinese Journal of Computers*(pp.1-14).

[3] Yinlei L, Yubao L, Cheng C.(2016).Efficient Mining Algorithm of Frequent Itemsets for Uncertain Data Streams. *Journal of Computer Research and Development*(pp.1-7).

[4] Mukherjee A, Kumar A, Bing L, Junhui W,Meichun H, Castellanos M, Ghosh R.(2013). Spotting opinion spammers using behavioral footprints. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*(pp.632-640). ACM.

[5] Haddadi H.(2010). Fighting online click-fraud using bluff ads.*ACM SIGCOMN Computer Communication Review*(pp.21-25).

[6] Shiguo C, Daoqiang Z.(2011). Experimental Comparisons of Semi-Supervised Dimensional Reduction Methods. *Journal of Software*(pp.28-43).

[7] Oentaryo R, Lim E P, Finegold M, et al.(2014). Detecting Click Fraud in Online Advertising: A Data Mining Approach.*Journal of Machine Learning Research*(pp. 99-140)

[8] Qian M, Ke Y.(2014).Overview of Web Spammer Detection. *Journal of software*(pp.1505-1526).

[9] Mukherjee A, Bing L, Glance N.(2012). Spotting fake reviewer groups in consumer reviews. *Proceeding of the 21st International Conference on World Wide Web*(pp. 191-200). ACM.

[10] Lee K, Caverlee J, Webb S.(2010). Uncovering social spammers: social honey-pots & machine learning. *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*(pp. 435-442). ACM.

[11] Metwally A, Paduano M.(2011). Estimating the number of users behind IP addresses for combating abusive traffic. *Proceeding of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp.249-257). ACM.

[12] Ravisankar P, Ravi V, Raghava Rao G, Bose I.(2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*(pp.491-500).

[13] Chaang C C, Lin C J.(2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent & Technology*(pp.389-396).

[14] Zibin D, Hangtian L.(2016). Efficient range matching method based on bloom filter and ternary content addressable memory. *Journal of Electronics & Information Technology*(pp.1872-1879).
[15] Wei L, Dafang Z, Kun H, Kun X.(2015). Accurate Multi-Dimension Counting Bloom Filter for Big Data Processing. *Acta Electronica Sinica*(pp.652-657).
[16] Dong C, Chen L, Wen Z.(2013). When private set intersection meets big data: an efficient and scalable protocol. *Acm Sigsac Conference on Computer & Communications Security*(pp.789-800). ACM.