Research on Data Mining Model in the Internet of Things

Yan Chen, Ai-xia Han Cai-hua Zhang

Department of Computer Technology, Hebei College of Industry and Technology, Shijiazhuang, 050000, China

Keywords: Data mining techniques; Crowdsourcing; Forged webs; Internet security; Internet of things

Abstract. The Internet of things combines aspects and technologies that come from several kinds of approaches. So the Internet of things is a new paradigm. When the computing, pervasive computing, and the relevant technologies are more closely interdependent to each other and the world of Internet of things form a system. More and more people become fond in the real and digital worlds, and continuously in symbiotic interaction. In this situation, to protect the online users such as the usernames and passwords is important. One intelligent approach based on data mining called Associative Classification seems a potential solution that may effectively detect phishing websites with high accuracy. We first review relationship between the data mining and the Internet of things. Then we highlight the potential solution. Finally, we conclude the potential solution is a effective way to protect the online users.

1 Introduction

The last ten years, mobile devices and ubiquitous services, that can give people a way to connect other people anytime-anywhere. The reason is the huge advancement in the field of electronics and the deployments of wireless communication systems. This moment, however, the role played by devices is no longer limited to connect the online user to the Internet of things. But, the devices have been expanding becoming an opportunity to interlink the physical world with the cyber world, in another word, the virtual world [1].

A uniquely continuum addressable things communicating, that can help one another to establish a worldwide dynamic network, form a new paradigm. The new paradigm has another name that is the Internet of the things. The partners of the auto-id center from the MIT is the founder of the the Internet of the things. We think the ecosystem of the country that is regarded as an case and we try do our best to find the answer to what the ecosystems of the country will look like, then these thing make the Internet of the things' define becomes more clear.

Not only individual users buy some goods online, the Internet of things. but also the organizations run their business online, the Internet of things. Many of the organizations set the stage, so the businessman can trade each other and the shopkeeper also can sale service and goods on the stage (Liu &Ye, 2001). So, a channel that is the commercial exchanges is the internet suitability, and the problem is about the internet suitability.

This paper is divided into different sections. Section 2 surveys same relationship with learning ways to data mining. The process of data mining and the model are illustrated in Section 3. Section 4 we discuss a real database case study. Lastly, conclusions are described in Section5.

2 Approaches to data mining and phishing detection

In this section, there are two important parts. One is the approaches to the data mining and the phishing classification. These are interpreted in the papers. The other talks about the phishing life cycle.

2.1 Data mining.

The center part in the whole data mining process is the mining phase. Classification, clustering, and association rule mining are the some kinds that data mining task could be divided. In the many field, the most classification is the face recognition, disaster rescue.

Many algorithms use "support confidence framework", such a structure can sometimes produce some erroneous results. Sometimes a rule's support and confidence than another implication of positive association rules is low, but it may be more accurate. If we take the support and confidence set low enough, then we will get two contradictory rules. On the other hand, if we get those parameters are set high enough, we can only get the imprecise rules.

In short, no one to support and confidence combination can produce association completely correct.

People through the study found that the interesting, can be used to prune uninteresting rules. In general, the rule interestingness is based on strength and expected real statistical independence assumption of the intensity ratio. However has been found in many applications, as long as people still regard support as the main decision set is initially produced factors. So, either to support low enough so as not to lose any meaningful rules, or take the risk of missing some important rules for the former case; the computational efficiency is not high, then a situation you may lose from the user's perspective is meaningful rules.

2.2 Approaches to minimize phishing.

In data mining associative classification is a promising approach, and can take use of special extraction from phishing and legitimate sites, finding (Costa Rica and Ortale, and 2013; Ritacco models Thabtah, fairing, Peng, 2005). Send an e-mail may be attacked by the phishing soft. It can be described that phishing attacks seem to be a real organization from the user. Via a link in an email, and ask them only to change their personal information.

Under normal circumstances, the two methods that are about the most technical in the fight against phishing attacks are belonging to the blacklisted and heuristic-based. Anti-phishing technology's success depends recognize phishing sites and deeply thinks about an acceptable period of time. Blacklist is considered to be malicious and collect those technologies already in use, such as the user's list of URLs vote. Another method is based on the comparison rules by collecting a series of Web sites different characteristics, the exposed classification algorithm. All the methods work, they need data mining.

3 Data mining process and the mining model

Pattern Discovery mode huge amount of data from a potential process is data mining. It is logical database system based envisaged. It is a multidisciplinary topic. The method of crowdsourcing can be used to complete various types of data mining classification tasks, such as clustering, semi-supervised learning, association rule mining.

The relationship of the database’s developing is depended on the data mining of the data base. For establishing the analysis systems to analysis the data set, the significant effort is find an efficient way. In general, the reason of the building the necessary data set is the collection of the usual tables database, that have to become coming together, be joined and transformed.

3.1 Classification process.

Basic work step is to tap the original data. The first step is to create a separator. The separator may define characteristics of the existing data. The second step is mainly used to classify. The users on the lines also can accomplish the categorization of the files. Custom user's front page news aggregator uses a Dig. Dig also allows users to submit links labeled, expand the scope to include more relevant articles of the user may be interested in, we found that the accuracy as more and more users get involved

3.2 The core algorithm

Agrawal designing is a basic algorithm in 1993, put forward an important method of association rules mining, which is a method based on the two stage frequency set of ideas, the association rules mining algorithm design is decomposed into two sub problems:

It finds all the support is greater than minimum support of itemsets, these itemsets called frequency set.

(2) using the first step to find the frequency set to produce the desired rule.

The second step here is relatively simple point. If a given frequency set $W = R_1 R_2 \ldots R_k, k \geq 2, R_j \in R$,Produce only contains all the rules set in the $\{R_1, R_2, \ldots, R_k\}$ .Each rule is only a right,once these rules are generated, then only those minimum confidence greater than the user specified rules to be left. In order to generate all frequent sets, it uses a recursive method.

Firstly generating frequent 1- itemsets M_1, then the frequent 2- itemsets M_2, until there is a f make Mf is empty, then the algorithm stops. Here in the article j cycles, process first to generate candidate itemsets B_j, B_j every itemset in only one out of two different belongs to R_{k-1} frequency set to do a $(J-2)$ connection to produce. B_k set in the frequency set is used to generate candidate sets, a subset of the final frequency set R_k must be B_k sets. B_k each element required for verification in the transaction database to determine whether the join R_k, the verification process here is a bottleneck of the algorithm performance. This method requires multiple scans may be large transaction database, this will increase a lot of I/O load.

3.3 The mining model.

We chose an algorithm, which is the algorithm is the ER model. We chose this model because the classification database changes. This algorithm can complete database schema and SQL queries to build and conversion, such as sequence databases.

Command S = {S1, S2,…,Sm}is a set of entities existing source said original table ER models. Please note that the source table is S, a collection, and conversion table is T1, T2, and so on. Note the use of the left outer table calculated on a query T0 universe. From the data mining perspective, we will establish a regression model as input, where the data set will have several explanatory variables and the target variable data sets. These variables do not exist in the database: they will change the query to obtain. Is a program of the method is as follows:

```
/* q0 : T0 , universe    */
SELECT I, /* I is the record id ,or point id mathematically */
        CASE
            WHEN D1 = ‘safety’ or D2 = ‘employed’ THEN 1
             ELSE
         END AS Y ,/*binary target variable */
INTO T0
FROM S1;
/* q1: deformalize and filter valid records      */
SELECT S2. I , S2. L, W3, W4, W5, W6, O1, O2
/* I is the record id, or point id mathematically */
        CASE
            WHEN
             ELSE
         END AS Y ,/*binary target variable */
INTO T0
FROM S1 JOIN S2 ON S1.I =S2.I
WHERE W5>10
/* q2 deformalize and filter valid records */
SELECT I,
        CASE
            WHEN B01 = ‘safety’ or B02 = ‘employed’
              THEN 00
              ELSE
         END
INTO T0
```

FROM S1;

Extend algorithm to a data mining model with transformation entities.

Input: the patterns are the S1, S2… and the query sequence is l_0, l_1, l_2 …

Output: R0, R1, R2 …

4 Case study

This section is mainly about the two issues. The first question is about the database and data mining problems. The second problem is that we analyze and interpret the data mining analysis dataset. The key point of the analysis is to explain the main features of each property and judgment. C # programming language is the algorithm we use. Microsoft SQL Server is selected articles database management system.

4.1 Input: Existing ER model and SQL script.

In this part, the most junior of the ER model and be converted, and SQL queries to the y0, y1... are we to explain. Then, the sequence generated by the model is a data set M. The final data set, the following, more detailed description of the properties and the end p = 16 n = 11,911 records to calculate a classification model. Through a statistical point of view, the variable value to our customers belong to the scope of the data set is multivariate statistics.

4.2 Output: New entities classified and extended ER model.

The output of our algorithm to generate entity names and attributes, as well as the classification of their type of conversion we first displayed. W3 summary is based on detailed information about the sales of the products purchased by the customer. Please note that the data mining model variables intended to explain why customers come back to the store. W5 produce the final data set O where the only record of each client. Polymerization to create the digital and categorical variables for each customer details archives. Please note that this client would not have a separate analysis summary for each product, but it is feasible to create a more accurate classification model, adding variables for each product or product category.

So if we use data mining models, functional phishing websites are sure to be found. Then, we use the results of the analysis can make appropriate measures. The goal is to teach users how to tell all online phishing sites and phishing scams soft. The ultimate goal - to ensure the safety of things, can be realized.

5 Conclusions

Internet social networking and online transactions daily confronted with the most important question is harmful phishing sites. In real life, this is because the fishing site is the network security issues cannot be ignored. In the near future, we consider the function-based content. The goal is to improve the collection of functions. We believe this will be potential research directions. The reason is that it helps to understand the attacker's behavior, which may help us to improve the performance of the method. Finally, we use data mining process and data mining cases, and by studying data mining method and found to be the selected method is effective, especially to protect online users.

References

[1] M.Conti, S.K.Das, C. Bisdikian, M. Kumar, L.M. Ni, A. Passarella, G. Roussos, G. Tröster, G. Tsudik, F. Zambonelli, Looking ahead in pervasive computing: challenges and opportunities in the era of cyber-physical convergence, Pervez. Mob. Comput. 8 (1) (2012) 2–21.

[2] K. Gama, L. Touseaui, D. Donsez, Combining heterogeneous service technologies for building an internet of things middleware, Comp.Commun. 35 (4) (2012) 405–417.

[3] C. Ordonez , Z. Chen, Horizontal aggregations in SQL to prepare data sets for data mining analysis, IEEE Trans. Knowl. Data Eng. 24 (4) (2012) 678 –691.

[4] S. Ceri, E. Valle, D. Pedreschi, R. Trasarti, Mega-modeling for big data analytics, Proc. ER Conference, 2012, pp. 1–15.
[5] T. Ku, Y. Zhu, K. Hu, A novel complex event mining network for monitoring RFID-enable application, in: Pacific–Asia Workshop on Computational Intelligence and Industrial Application, 2008. PACIIA ’08, 2008, pp. 925–929.
[6] A. Dias, L. Gorzelniak, R.A. Jrres, R. Fischer, G. Hartvigsen, A. Horsch, Assessing physical activity in the daily life of cystic fibrosis patients, Pervas. Mob.Comput. 8 (6) (2012) 837–844.
[7] TAN Ying. A prototype architecture for cyber physical system s[J] . SIGBED Review,2008 ,5(1) : 51-52.
[8] J.Kennedy, RC.Eberhart. A discrete binary version of the particle swarm optimization algorithm. Proceeding of International Conference on System, Man, and Cybernetics, 1997: 4104-4109.
[9] Yoshida H., Fukuama Y., Takayama S. A Particle Swarm Optunization for Reactive Power and Voltage Control in Electric Power Systems Considering Voltage Security Assessment. IEEE International Conference on Systems, Man, and Cybernetics. 1999, 6(497): 502.