Differentiated security levels for personal identifiable information in identity management system

Jianyong Chen a,*, Guihua Wu a, Linlin Shen a,b, Zhen Ji a

a Department of Computer Science and Technology, Shenzhen University, PR China
b State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, PR China

Abstract

With the rapid development of Internet services, identity management (IdM) has got widely attraction as the credit agency between users and service providers. It facilitates users to use the Internet service, promotes service providers to enrich services, and makes Internet more security. Personally identifiable information (PII) is the most important information asset with which identity provider (IdP) can provide various services. Since PII is sensitive to users, it has become a serious problem that PII is leaked, illegal selected, illegal accessed. In order to improve security of PII, this study develops a novel framework using data mining to forecast information asset value and find appropriate security level for protecting user PII. The framework has two stages. In the first stage, user information asset is forecasted by data mining tool (decision tree) from PII database. Then security level for user PII is determined by the information asset value assuming that the higher information asset is, the more security requirement of PII is. In the second stage, with time being, number of illegal access and attack can be accumulated. It can be used to reconstruct the decision tree and update the knowledge base combined with the result of the first stage. Thus security level of PII can be timely adjusted and the protection of PII can be guaranteed even when security threat changes. Furthermore, an empirical case was studied in a user dataset to demonstrate the protection decision derived from the framework for various PII. Simulation results show that the framework with data mining can protect PII effectively. Our work can benefit the development of e-business service.

1. Introduction

With development of information processing and electronic commerce, more and more people use Internet to communicate with each other and do shopping. Identity management (IdM) (Thompson & Thompson, 2007) is the most important platform for various application services on Internet. It enables all entities to interactive each other successfully on Internet. IdM is concerned with identities that can identify individuals uniquely within a given environment, which is designed to store, protect and manage user personally identifiable information (PII). Recently, IdM has got widely attraction as the important means to set up credit between users and service providers (SP). It can facilitate users to experience the Internet service, benefit service providers to reduce investment on services, and make Internet more security. IdM is composed with three types of fundamental entities with User, identity provider (IdP) and SP. Among them, IdP is the core entity to perform IdM functions. SP is responsible to provide application service. Normally, when user requests service from SP, SP sends query to IdP to ask for user identity resource. SP will provide user the service if user identity resource satisfies SP’s requirement.

1.1. Protection of PII

PII refers to natural person’s special attributes that can directly or indirectly identify who they are. It could include user’s name, social identity (ID), preferences, address and other information. Since PII involves user privacy and sensitive information, the protection of PII is one of key issues in IdM. IdP needs enough hardware and software to ensure the security of PII. It becomes a serious problem that PII is leaked, illegal selected, illegal accessed (Mont, 2004). Many international organizations and individuals, such as IBM and Microsoft, have put much resource to the research and development (R&D) of protection of PII. Many frameworks and security technologies are proposed to protect user PII. For examples, Security assertion markup language (SAML) (Cantor, Kemp, Philpott, & Maler, 2005) and Ws-security (WS-Federation, 2006; WS-Trust Specification, 2007) protocols are developed to provide secure communication between IdPs and SP. With digital signature and encryption technologies (Acquisti, 2008), mechanisms are defined to create assertions as security tokens which can be used...
to protect user identity information. However, it is not necessary for all applications to use the same strength of security since the improvement of security strength means additional consumption of IT resource. Especially, in cloud computing scenarios, differentiated security level for user PII is important for the platform to provide high cost performance service to all users. However, there are not effective methods to apply the differentiated security on protection of PII. This study aims to develop a data mining framework for IdM system to automatically predict and set security levels for PII using useful patterns and rules that explored from PII data. When user registers in an IdP by applying the proposed framework, PII can be classified to be protected according to security levels. A case study on PII dataset is studied to demonstrate the validity of this approach. Simulation results show that the framework with data mining can protect PII effectively and perform better than IdM system without data mining.

1.2. Paper organization

The rest of the paper is organized as follows: In Section 2, differentiated security for PII is introduced. In Section 3, the framework using data mining tool is proposed and the corresponding process and components are discussed in detail. Data preparation is described in Section 4. The construction of decision tree is represented in Section 5. In Section 6 the performance of the framework using data mining is analyzed and compared with IdM system without data mining. Section 7 describes the protection of user PII according to differentiated security policies. In Section 8, considering attackers' character, the decision tree for differentiated security is reconstructed and the security level of PII based on times of attack is set up. Section 9 is conclusion of this paper.

2. Differentiated security for protecting personal PII

Security requirements of PII vary dramatically among different users that are stored and transmitted on network. In networking applications, communication is important, such as e-commerce, etc. But some applications just require low level of security, such as accessing to Internet for open information. Comparing with unsecured information, the secure information means additional investment from the SP's point of view. Moreover, there is no absolute security of information. The investment on security strongly relies on the level of security the system can provide. The additional investment includes network management about security, security devices, additional consumption on bandwidth and computing power, the training for security managers and users, etc. Evidently, the structure of security policy is necessary to provide the differentiated security (Chen, Wang, & He, 2008). According to the identity and context of users or resources, differentiated security uses different security policies and mechanisms to protect user information or resources information. In this case, it is more difficult to scale or replicate attacks, since each individual or each class has a different security profile and there should be no common weaknesses. By this means, it is important to study the differentiated security. In user-centric IdM systems (Marit, Ari, & Alissa, 2008), security policies are made according to user preferences. IdM systems support selective mechanisms by giving user full control on what circumstances and which identity attributes are shared. In the standards of IdM system, there has been a clear definition on security levels of protection of PII (ITU-T Recommendation SG17 TD0130, 2009). Therefore, the value of user PII is important to determine the security levels of PII.

According to differentiated security, the protection of PII is divided into the following steps:

(a) Maintain user PII. In this phase, through user registers in IdP, user PII is collected and selected for user identity service.
(b) Evaluate the value of user PII. From users' perspective, their identity information is important and is seen as their asset, namely, information asset. Its value of information asset determines the security level of protection of PII. A method to evaluate the value of information asset is performed systematically and effectively. From the viewpoint of enterprise benefit, since the security level of protection of PII in IdM system are determined by the value of the asset, the evaluation of information asset is an important activity.
(c) After the information asset value is evaluated according to the method established by IdM system, the security level is decided. A security level of protection of PII is used according to the value of information asset.

3. Framework for protecting user PII

3.1. Main factors on security level

Security level of user PII is determined by both information asset and attacking frequency on the PII. Since it is impossible to measure attacking frequency just after user registration, it is important to evaluate user information asset at the initial stage. With the time being, the measurement of attacking frequency on the PII can be accumulated and is used to optimize the security level dynamically. The security level will continuously be updated from the combination of the initial security level and the attacking frequency. Fig. 1 shows factors that can affect security level for user PII. In this figure, information asset is evaluated from various attributes. It determines the initial security level of user PII. After the user PII is registered, the user PII may suffer from various attacking or illegal accessing during its lifecycle. The attacking frequency is measured which as well as the initial security level is used to determine the current security level dynamically.

3.2. Framework

In this section, a framework is represented to make decision on protection of PII. The framework has two processes stages. In the first stage, security level for user PII is determined by the value of information asset. In the process of establishment of decision tree, data is trained according to value of the asset. With user information asset as classification label, the framework builds a decision tree to forecast the value of information asset from a new registration. Then corresponding security level is made to protect the user PII. In the second stage, with time being, the number of illegal accessing and attacking to PII can be accumulated. It can be used to reconstruct the decision tree and update the knowledge.
The proposed framework enables IdP to forecast user information asset, the underlying threat and to find appropriate security level for protection of PII. This framework includes four major parts: PII database, data mining tool, knowledge base, and intrusion detection, as shown in Fig. 2. Their functions are presented as follows.

(i) PII database
IdP is responsible for user registration, identity validation, identity authentication and management of PII. It is the core entity to establish trust relationship between SP and the user of an IdM system. In identity register phase, user PII including user contact and other privacies is committed to IdP. In the remains lifecycle of user identity, PII is stored in database and security measures are adopted to protect it.

(ii) Data mining tool
Data mining is used to extract underlying information and prediction of patterns. Data mining includes association rule (Chen & Weng, 2009), clustering (Perera, Kay, Koprinska, et al., 2009) and classification (Mangasarian & Wild, 2008). It has been applied in many domains and industries successfully, such as customer relationship management (Ngai, Xiu, & Chau, 2009), identification of disease (Yang, Zhao, Zhu, et al., 2008), production operation sequence in industry (Rokach, Romano, & Maimon, 2008), medical decision making (Speckauskiene & Lukosevicius, 2009), human resource management (Zhao, 2008), knowledge management (Cunha, Adeodato, & Meira, 2009), industry standards development (Chih-Hung, 2009), social welfare decision (Chin-Jui & Shiahn-Wern, 2009) and patent analysis (Yan-Ru, Leuo-Hong, & Chao-Fu, 2009).

Various classification algorithms can be used in this framework as data mining tool. In this paper, decision tree classification algorithm, i.e., C4.5, is used to extract underlying information. It is a typical algorithm used to generate a decision tree developed by (Quinlan, 1993). C4.5 constructs a decision tree based on training data. The training data is a classified sample set \( S = \{s_1, s_2, \ldots, s_n\} \). Each sample data \( s_i = (attr_1, attr_2, \ldots, attr_n, c_i) \) is a vector in which \( attr_1, attr_2, \ldots, attr_n \) represent attributes of the sample and \( c_i \in C \) represents the class of user record. The vector \( C = (c_1, c_2, \ldots, c_n) \) includes the class labels: \( c_1, c_2, \ldots, c_n \). A class label is the class that one record of training data belongs to. After test on the decision tree, the patterns included in the decision tree are used to forecast the PII security of new register from IdM system. The patterns may facilitate IdM system to manage user PII. Decision tree can deal with general and numeric data. It can provide a good decision on large data resource in a short time. And it is easy to be understood and implemented.

(iii) Knowledge base
Results from decision tree are translated to rules and knowledge that are used to determine security levels for the protection of PII, which is stored in knowledge base. If knowledge base is built, security level of particular user PII can be got and is further used to determine the PII of new register users with the same classification of PII.

(iv) Intrusion detection
Intrusion detection is used to record times that user PII is attacked. In the second process, with time being, number of illegal accessing and attacking can be accumulated by intrusion detection. It can be used to reconstruct the decision tree and update the knowledge base. The forecasting result will be more precise and the security level of PII can be timely adjusted. Thus the protection of PII can be guaranteed even when security threat changes. In this framework, the four described modules interact with each other to improve the security of PII. On the one hand, data mining tool is used in PII database to extract useful information and patterns to construct decision tree. The useful information and patterns are stored in knowledge base. On the other hand, the number of attacking that PII database suffered is also accumulated and recorded by intrusion detection. It is used in second stage to reconstruct decision tree and dynamically forecast the security level that users PII need. Thus, security of user PII is ensured.

3.3. The process of framework
The framework aims at providing IdM system with protecting suggestion on PII. Associating with data mining, the process of this

![Fig. 3. The process flow of framework.](image-url)
framework in first stage includes those procedures as shown in Fig. 3. The process concerns the data mining on application domain. It is composed by the following sub-process:

- Target definition: IdP in an IdM system is responsible for management of user PII, whose target is to ensure the security of PII in its lifestyle, including the use, transmission and store of PII. Different kinds of users need use specific security level. In this application, the target also includes the determination of security level for PII from new register user.
- Data preparation: User PII data is extracted from PII database in IdP and is preprocessed.
- Data mining: Data mining includes data discretization, handling of noise data, processing of data with fragmentary attributes and the generation of decision tree. Decision tree algorithm is operated to extract useful information and patterns.
- Decision tree evaluation: Evaluation methods are used to validate its effect where untrained data is also used to test the decision tree.
- Application of data mining result: The result acquired from decision tree is translated to knowledge and rules. Then the rules and knowledge is used to manage user PII in IdM system.

At first stage, the security level for protecting user PII is established. But the protection decision is determined just by information asset value from the perspective of user and IdM system. Attackers' preferences are not reflected in the result. In the second stage, attackers' preferences which are used as class label to reconstruct decision tree are taken into consideration. At last, the results from two stages are combined to compute final security level for user PII.

4. Data preparation

Generally, security level for protecting PII is determined by user attributes which refers to user privacy, such as user address and user contact. From the security perspective, the security requirement of attributes is different according to user preferences and privacy request. The security level for protecting PII need be determined by the most important attribute. “adult” dataset from UCI (http://www.ics.uci.edu/~mlearn) including 32561 instances is selected as user register information and user information asset is selected as classification label of decision tree algorithm. Users come from United-States, Cambodia, England, Puerto-Rico and etc. They work at eight different places, such as private company, federal government, local government and etc. Their education covers various levels. Users whose information asset is above 50K and the users whose information asset is less than or equal to 50K are 31.72% and 68.28%, respectively. The variables used for the data mining are presented in Table 1. The data mining algorithm is applied to find which attributes are significant at prediction of information asset value. In this paper, the results of data mining for the first stage are presented.

5. Decision tree construction

5.1. Overview

Decision tree is constructed by user dataset and tested by test data. User’s data is comprised of various attributes. In this experiment, the datasets are appropriate to show the effectiveness of the proposed framework in this paper. In the process of construction of decision tree, 10-fold cross-validation is used as test model to test the tree. The result shows that since user information asset correlates with multiple attributes, the proposed framework with data mining can get more accurate secure level of user PII than classification with single attribute without data mining tools.

5.2. Decision tree construction

WEKA (Waikato Environment for Knowledge Analysis) (University of Waikato) software is developed by Waikato University, New Zealand. WEKA as knowledge discovery software has integrated many machine learning tools in which C4.5 is available as J48 classifier tree.

According to the IdM system without data mining, information asset value is forecasted only by attribute “Workclass”. If class of work is “Self-emp-inc”, the information asset value is above 50K. When class of work is others, the information asset value is less or equal to 50K, such as “Private”, “Self-emp-inc” and “Federal-gov”. At last, the instances in dataset that are correctly classified into their classes are 24818 and the corresponding percentage is 76.22%. Classification result on user information asset value without data mining tool is represented in Fig. 4. User information asset value is predicted only based on attribute “workclass”. In this case, it is not enough accurate and unfair to all of the users PII. Table 2 represents true distribution of user information asset value and the wrong classified instances corresponding to attribute values of “workclass”. If IdM system only uses attribute “workclass” to forecast user information asset value, the total instances classified into wrong classes are 7734. Here, classification is wrong with the following attributes: 4963 instances with attribute value “Private”, 724 instances with attribute value “Self-emp-not-inc”, 494 instances with attribute value “Self-emp-inc”. Those wrong classifications increase the security violation to user PII. The granular is too coarse if user information asset value is only determined by single attribute.

In the proposed framework, IdM system with data mining tool is used to forecast user information asset value. The decision tree is built by WEAK, including 607 nodes with 485 leaf nodes. It means that there are 485 paths that are composed of different attributes to classify users’ register information. It is smaller granular than that in IdM system without data mining. The tree is big for that each attribute has many attribute values. However, when the test model is used with 10-fold cross validation, the percentage

<table>
<thead>
<tr>
<th>Attribute name</th>
<th>Attribute description</th>
<th>Attribute values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>User age</td>
<td>Real type</td>
</tr>
<tr>
<td>Workclass</td>
<td>The class of user</td>
<td>Private, Self-emp-not-inc, Self-emp-inc, Federal-gov and etc.</td>
</tr>
<tr>
<td>Education</td>
<td>User education level</td>
<td>Bachelors, Some-college, 11th, and etc.</td>
</tr>
<tr>
<td>Education-num</td>
<td>Number of user</td>
<td>Real type</td>
</tr>
<tr>
<td>education year</td>
<td>education year</td>
<td></td>
</tr>
<tr>
<td>Marital-status</td>
<td>user marriage status</td>
<td>Married-civ-spouse, Divorced, Never-married, etc.</td>
</tr>
<tr>
<td>Occupation</td>
<td>User occupation</td>
<td>Tech-support, Craft-repair, Other-service, Sales, and etc.</td>
</tr>
<tr>
<td>Relationship</td>
<td>User relationship in</td>
<td>Wife, Own-child, Husband, etc.</td>
</tr>
<tr>
<td>family</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td>User race</td>
<td>White, Black Asian-Pac-Islander, etc.</td>
</tr>
<tr>
<td>Sex</td>
<td>User sex</td>
<td>Female, smale</td>
</tr>
<tr>
<td>Capital-gain</td>
<td>User capital gain</td>
<td>Real type</td>
</tr>
<tr>
<td>Capital-loss</td>
<td>User capital loss</td>
<td>Real type</td>
</tr>
<tr>
<td>Hours-per-week</td>
<td>user work time in a</td>
<td>Real types</td>
</tr>
<tr>
<td>week</td>
<td>week</td>
<td></td>
</tr>
<tr>
<td>Native-country</td>
<td>User country</td>
<td>United-States, Cambodia, England, Puerto-Rico, etc.</td>
</tr>
<tr>
<td>Information asset</td>
<td>User information asset</td>
<td>&lt;=50K, &gt;50K</td>
</tr>
</tbody>
</table>
is about 86.26% for the correct classified instances. The decision tree is shown in Figs. 5 and 6. Seen from the figures, attribute "capital-gain" is important to find user information asset which is the first classification boundary. When attribute value of "capital-gain" is less than or equal to 6849, marital-status, capital-loss, education-num and hours-per-week are important to user information asset value as shown in Fig. 5. In (1) attributes "Marital-status", "Capital-loss", "Education-num", "hours-per-week", and "relationship" are used to forecast user information asset value sequentially. In (2) attributes "Marital-status", "Capital-loss", "Education-num", "Hours-per-week", "Age", "Sex" and "Occupation" are used to forecast user information asset value sequentially. In (3) attributes "Marital-status", "Capital-loss", "Education-num", "Hours-per-week", "Age", "Occupation", and "Workclass" are used to forecast user information asset value sequentially. On the other hand, when the attribute value of "Capital-gain" is above 6849, "Education-num", "Hours-per-week", user "Age" and "Education" are vital to user information asset as shown in Fig. 6. According to different values of attribute used to classification, the tree is divided into sub-trees. It is evident that user information asset value is related to many attributes. It is necessary to extract the patterns with data mining technology. And the deployment of security levels for PII becomes reasonable and evidently different from the current IdM system without data mining.

6. Performance evaluation

In evaluation phase, receive operating characteristic (ROC) Tom, 2006 curve is used to visualize and analyze classifier's performance. An ROC graph is a technique for visualizing, organizing and selecting classifiers based on their performance. In ROC graph, if ROC curve is beeline with $y = x$, it means that the corresponding classifier is the worst and it can not be used for any classification questions.

Using ROC curve plot function of WEAK, the evaluation result of IdM system with and without data mining is drawn in Figs. 7 and 8. In IdM system without data mining, both "Workclass" and "Education" are used to forecast user information asset value. Fig. 6 represents the effects using three different methods when user information asset class label is above 50K. X-axis represents proportion of correctly classified instances in the instances which belong to class label "information asset > 50K". Y-axis represents proportion of wrong classified instances in the instances which belong to class "information asset <= 50K". Fig. 7 represents the effects using three different methods when user information asset class label is equal or less than 50K.

In IdM system without data mining, whatever in the classification of information asset ">50K" or "<=50K" when "Workclass" and "Education" are used to forecast user information asset, both the ROC curves are closing to beeline with $y = x$, especially for "Workclass". But the ROC curve with decision tree is away from $y = x$. Evidently, it is more accurate for IdM system to use decision tree to predict the user information asset than that without data mining. Furthermore, Area Under the ROC Curve (AUC) Huang and Ling (2005) can illuminate their difference in performance accurately. Their AUC is showed in Table 3.

According to the definitions of ROC and AUC, it is evidently that the AUC value of an ideal classifier is equal to 1. Seen from Table 2, whatever information asset value is above 50K or less than or equal to 50K, its AUC is 0.894. It indicates precisely that proposed frame-
work has good performance. When the decision tree is established, a test dataset is used to test it. The test dataset has 16281 instances and 13992 instances are correctly classified, whose percentage is 85.94%. The test result proves that the decision tree is efficient to forecast new user PII.

7. Application

Results from decision tree algorithm are stored in Knowledge base. IdP can get security level of individual user PII from the knowledge base. When user fills the necessary identity information and registers in IdP, user’s security level is found from knowledge base. In this paper, security levels are divided into two levels: 1 and 2, however, security level can be different in varies of applications. Fig. 9 describes a part of the results in the knowledge base.

As seen from the rules, it is easy to check whether user’s PII would be attacked and what security level should be adopted. When user’s capital-gain is 7000, education-num is 9, hours-per-week is 30 and age is 23, according to rule 2.1, security level of user’s PII is 1. If capital-gain is 8000, education-num is 8, hours-per-week is 28 and age is 31, according to rule 2.2, the security level of user’s PII is 2.

8. Reconstruct decision tree

In the first stage, IdP has no intrusion detection records about users, user information asset is used as class label for decision tree. With time being, number of illegal access and attack of PII is accumulated by intrusion detector. In the second stage, the number of attacking is used to reconstruct decision tree and update the knowledge base. After decision tree being reconstructed, the attack model of user PII is mined easily by using both the times of attacking and information asset value because the times of attacking can be detected. The main process of reconstructing the decision tree is described as follow:

- Times of attacking on each user are added into user data as another class label. The updated dataset is used for reconstruct decision tree.
- C4.5 is used to build decision tree based on times of attacking. The process is the same with the initial stage.
- Security level based on times of attacking. According to times of attacking, security level is set to protect user PII.
- Update the security level for protecting PII.

As aforementioned reason, user information asset is important facet which may introduce security violation. However, times of attacking also reflect the fact that users are attacked in networking. For the reasons, in second stage, they are combined to determine the final security level as the following equation:

$$I = \max(I_1, I_2)$$

In Eq. (1), $I_1$ is the security level found based on the information asset forecasted by decision tree in initial stage. $I_2$ is the security level found based on the times of attacking forecasted by decision tree in second stage. And $I$ is the final security level used for IdM system to protect user PII. According to default worst-case scenario in the principle of risk management (Barry, 1991), the security level of the framework is able to prevent the worst security...
violation. So the final security level is the max value between $l_1$ and $l_2$. Thus user PII in IdM system always can acquire enough security protection.

9. Conclusions

This study applies a data mining framework with knowledge on IdM system to obtain appropriate security levels for user PII. Using the framework, IdM system can classify PII into appropriate security levels accurately. According to information asset value and assigned security level, IdM system can use suitable security mechanisms, such as encryption, to control the access of PII and avoid the leakage of PII. At the same time, with time going, the rules with decision tree can be reconstructed and the knowledge base can be updated automatically according to frequency that user PII is illegal accessed and attacked. This research can evidently improve the security of IdM system on user PII and thus can benefit the interoperation of IdPs in different trust circumstance.

Acknowledgements

The work was supported by China-Finland Cooperation Project on the Development and Demonstration of Intelligent Design Platform Driven by Living Lab Methodology (2010DFA12780).

References


http://www.ics.uci.edu/~learn


