Insider Threat Detection and Prevention Protocol: ITDP

https://doi.org/10.3991/ijoe.v17i02.18297

Amnat Sawatnatee
Chandrakasem Rajaphat University, Bangkok, Thailand

Somchai Prakancharoen (✉)
King Mongkut's University of Technology North Bangkok,
Bangkok, Thailand
somchai-prakan@hotmail.com

Abstract—Insider threat is a severe problem for many computer departments since they have an authorization to do some assigned tasks. They can easily seek security for any organizational computer vulnerability. Protocol "Insider Threat Detection and Prevention Protocol: ITDP" is designed to detect whether a requesting "IT user" is an authentic IT user who has been allocated rights to a particular application. The User's knowledge and behavior are used to classify whether the user is authentic. The statistical classification technique is used to predict whether the guest is authentic. The best classification technique is linear binary discriminant function analysis with 98.3% of accuracy in insider threat detection classification.

Keywords—Insider threat, question-answering, computer usage behavior, rough set, binary logistic regression

1 Introduction

The security breach in the organizational data processing system has arisen from both external and internal intruders. Insider threat, who deceives another authentic "IT user", is an incident that is very difficult to prevent. The external attack can be detected and prevented by many mechanisms before they can enter the computer system. On the other hand, insider threats can easily be malicious seeking the key of some target "IT user". After that, he can get access to some application program to gain some profit or even to malign someone. This paper presents a practical Insider Threat Detection and Prevention Protocol: ITDP. All insider clients, "IT users", have to answer some questions besides their jobs; such as favorite food, dish, etc. Their answers were kept in the database for their future verification. Moreover, behavior of all "IT users" about start working time, stop working time, amount of working time and favorite website visiting is collected from many related log databases. All of these features are carefully used to consider if he is an authentic or fake "IT user". A Rough Set technique was used to select essential attributes and consistent behavior patterns. The calculated patterns were used to detect a cluster of "IT users" who have similar behavior. Someone else that is a

member of the same group of other "IT users" might easily get access to other's responsibility by assuming his name. This kind of "IT user" must be carefully detected by designed protocol before the system should allow them access to some application programs. ITDP offers a classification equation to the application administrator to identify whether the "IT user" is an authenticated "IT user".

2 Related Theory and Research

2.1 Rough set [1]

Rough Set theory is a mathematical tool that could discover data patterns from data analysis. It is used for decision rule extraction, feature extraction, data reduction and association rule. Indecision rule extraction, a special characteristic of Rough Set theory is that it can discover certain and uncertain decision rules. There are two types of attributes: conditional attribute (set A) and class or decision attribute (set D). Let IS (Information System) is a set of U and A. U is a nonempty finite set. A is a set of attributes $\{a_i\}$. $IS = (U, A)$. Each observation (set X, $x \in X$) is composed of attributes "a" ($a \in A$) and "D". This set of observations is called decision system or table (T). $T = (U, A \cup \{d\}); d \notin A$. Let $B \subseteq A$, B indiscernible (same) of any two observations (x, x') could be obtained based on the logical sentence as shown $IND_{IS}(B) = \{(x, x') \in U^2 \mid \forall a \in B, a(x) = a(x')\}$. This equivalence class based on "B" denoted as $[X]_B$. Inset of an equivalence class, if all "B" in the equivalence class $[X]_B$ is an element of "X" then the approximate "X" is called B-lower, denoted as $\underline{B}X = \{x \mid [X]_B \subseteq X\}$. If some of "B" in the equivalence class $[X]_B$ are an element of "X" then the approximate "X" is called B-upper, denoted as $\overline{BX} = \{x \mid [X]_B \cap X \neq \phi\}$. Accuracy of approximate can be calculated from the proportion of B-lower and B upper, $\alpha_B = \frac{\underline{B}X}{\overline{BX}}$. If its value is "1" then the approximation is "crisp" to "B". Elsewhere, $0 \leq \alpha_B \leq 1$ then "X" is "rough" to "B". Based on the decision rule, the Rough Set could consider if some conditional attribute is essential to keep a crisp or certain rule. In any case, some attributes could be ignored since it is not needed in crisp rule generation. The set of conditional attributes that are needed in rule generation are called "Reduct"

2.2 Discriminant Analysis [2]

Discriminant Analysis is a statistical technique used to classify observations into non-overlapping groups based on scores on one or more quantitative predictor variables. Each observation is assigned to a particular cluster based on its Discriminant value distance from the cluster's centroid. Discriminant function is calculated with the same

method of linear regression. The difference between the two approaches is that the Discriminant function dependent variable data type is a categorical variable.

2.3 Cryptography [3]

Information security goals are covered in secrecy (confidentiality), integrity and availability. Cryptography is a mathematical algorithm that could keep confidentiality and integrity. Symmetric or conventional key encryption, such as DES is fruitfully used in secrecy preservation. Whereas, Public key encryption cryptography, such as RSA, offers both secrecy and integrity. Public key encryption cryptography has two inverse keys. These two keys are generated by the key owner, such as "A". The first is called a public key, $K_{Pub\text{-}A}$. A public key is mostly used by his participant, such as "B". Normally, the public key will be given to someone that the key owner wants to communicate with. The second key is a private key, $K_{priv\text{-}A}$. The private key is kept secret by the owner. This key is used to represent his authentication. For example, if "A" wants to send a message "M" to "B" under the secrecy of sending the message and present of "A", an authentic message. Step 1. "A" performs cipher text: $E_{kpub-B}(E_{kpriv-A}(M))$ Step 2. "B" performs cipher decoding $D_{kpub-A}(D_{kpriv-B}(E_{kpub-B}(E_{kpriv-A}(M))))$.

Certification Authority (CA) is a third-party organization that takes the responsibility of a digital certificate issued to someone who registers to CA as a member. He has to send his public key and some formal identity, such as his ID card to CA. After cross-checking of formal identity, CA will append the applicant's public key in the CA database based on some protocol such as X.509. CA's member is then certified his authentication to his participant under his digital certification. When someone else, such as "B", wants to communicate with someone, such as "A" who is a CA's member, then "B" will ask "A" public key from CA. After that, "B" will communicate with "A" under message encryption with an "A" public key. Therefore, if "A" is not CS's member then "B" may gain risk in unsecured communication with "A".

2.4 Questioning technique [4]

Benjamin S. Bloom presented that human being's learning is covered in three types as cognitive domain, affective domain and psychomotor domain. Bloom's taxonomy is composed of six levels as knowledge, comprehension, application, analysis, synthesis and evaluation. Bloom's taxonomy is used to discriminate the level of learning. The teachers could measure their students' progressive learning by asking the various level type of questions. For example, there are many types of questions such as managerial questions, rhetorical questions, closed questions, open-ended questions. Generally, the same question type on some levels of learning of each student should have different answers since they always have a different way of life and educational foundation.

2.5 Insider threat [5]

Human behavioral factors of an organization employee that encourage insider security threats are grouped into many topics such as organizational weak security policy, regulation, practicum, employees under job evaluation, cyber loafing, financial concern, criminal record, ideology, etc. These conditional attributes were used to classify insider threat ontology. Nevertheless, some employees have an undesired attribute but is not an insider threat. Therefore, organizational experts or employers have to carefully observe and discriminate against this kind of employees.

2.6 Web usage mining [6]

Regularly computer system users have logged on to some web servers to get access to some servers' application or even connect to some websites. These activities are kept in server log-files, application server log, and web-log. Web usage mining is a technique used to discover the knowledge of IT user's behavior in computer system usage. This insights pattern could be used to enhance computing service performance. Moreover, each web usage pattern could be used to identify an "IT user" whether he works in normal operation or deception operation.

2.7 Related research

A Bayesian network model for predicting insider threats [7]: Malicious insider incentive and psychological conditional attributes were collected from much-related research. These gathered attributes were considered their critical importance or correlation on insider deception. Structural equation modeling was used to exploratory and confirmation conditional factors related to a class factor (malicious insider). After that, this empirical structural equation model was adjusted to be a Bayesian network model for predicting insider threats.

Modeling and verification of insider threats using logical analysis [8]: Florian et al have studied sociological explanations of organization infrastructure. The result of the study could explain conditional attributes that affect a class variable (insider threat). The study was specified on both normal and fake IT users. Observation data were transformed into formal modeling by using higher-order logic. Patterns of insider threats were summarized as insider threat theory.

An approach for intent identification by building on deception detection [9]: Based on past research in deception detection at the University of Arizona, the research result has guided to investigate intent detection. A theoretical foundation and model for the analysis of intent detection is proposed. Available testbeds for intent analysis are discussed and two proof-of-concept studies exploring nonverbal communication within the context of deception detection and intent analysis are shared. This research could present some techniques to find deception occurring.

End-to-end privacy protection for a Facebook mobile chat-based on AES with multi-layered MD5 [10]: Social media, such as Facebook is a popular social media in the world. It supports user's communication with their community. Chat is the most

favorite feature in its activities. Facebook always asks for the user's information. This information is used to connect each user to his friend of the friend. Unfortunately, the user's personal information may become a precious commodity. User's goods buying behavior in the market place depends on platform. Therefore, the secrecy of communication messages should be kept secret from both third-party and especially social media platforms. Wibisono [10] suggest private chat protocol between social media users by encrypting those messages with AES symmetric block cryptographic algorithm. The ciphertext is then hashed with a multilayered MD5 hashing function for integrity verification.

Cloud–internet communication security framework for the internet of smart devices [11]: Since internet communication speed is tremendously increasing, then the "Iot" has been rapidly developed. The internet of smart device networks is composed of sensors, wi-fi, communication frameworks and cloud system. Data storage and data processing are managed by cloud storage and cloud computing. Most security breaches occur while smart devices sending or receiving a message from itself with a cloud system via networking. Tanweer et al have developed a secure communication framework that could increase user's message secrecy and privacy between the internet smart device and cloud system.

A Novel authentication mechanism to prevent unauthorized service access to a mobile device in a distributed network [12]: The client-server is the distributed computer network architecture that client or user has to log on to the server for data processing. The server has to detect if the current log on client user is legitimate. Pavani suggests a security mechanism that could detect log on user client authenticated by RSA public-key cryptography, once he is logging on. After that, this client could securely connect to other computer resources by Diffy-Hellman, public-key system, session keys. The proposed mechanism could keep legitimate log on and give users comfortable on travel to other distributes computer network's resources.

Intensive pre-processing of KDD cup 99 for network intrusion classification using machine learning techniques [13]: A network security breach is an essential task that a computer network firewall has to detect and prevent. The signature of each intruder must be prior learned from a real intruder data package. Gathered Network intruder's observation from the KDD dataset was used to train for each intruder signature. Ibrahim found that the classification technique Random Forest Classifier gave more accuracy in classification than Random Tree, J-48, Naïve Bayes. However, data training has to frequently re-calculated since there are many new emergence intruders.

Integration of user profile in the search process according to the Bayesian approach [14]: An information retrieval technique is used to retrieve some information based on its related features. Farida suggests that the user's personalization profile is an important feature that could relate to their interest class variable. The Bayesian network was used to build a model of a classifier user profile with their interest information.

3 Insider Threat Detection and Prevention Protocol: ITDP Design

The ITDP protocol is designed to support computer usage operation of IT users or clients about data processing with some software applications. The stakeholder of this context composed of "IT user" or client, application security bot, log on-off administrator bot, and CA bot.

IT user log on into a computer system to get access to his/her obligated application program. If he/she has passed "password checking" then he/she can do any task as he/she has a pre-assigned application. It is a worse situation than someone who knows another one's password. ITDP suggests that each user has to register himself with CA to certify his authenticity under the public-key system as shown in figure 1, step 0.0, 0.1, 0.2, 1 and 2.

However, some intelligent insider intruders might gain someone public and private key thus prior tasks are not believable. ITDP offers an "Insider Deception Detection Module: IDDM" to manage IT user verification. Overall ITDP operation is explained in (A) and IDDM in (B).

a) Insider Threat Detection and Prevention Protocol: ITDP

An ITDP is composed of 12 tasks (3-14) to complete IT user's authenticity checking. While IDDM has responsibility in four tasks that directly relate to deception detection.

1. "Log-on user (p-1 application user)" log on to "IT administrator". $((\text{Emp\_ID})_{\text{Emp\_id}_{\text{K\_priv}}})_{\text{IT\_ad}_{\text{k\_pub}}}$
2. "IT administrator (IT_ad) informs "welcome" to "p-1 application user". $((\text{"Welcome"})_{\text{Emp\_id}_{\text{K\_pub}}})_{\text{IT\_ad}_{\text{k\_priv}}}$
3. "P-1 application user" request for P-1 application using to "P-1 security agent". $(\text{"}p-1\text{"}, Emp\_id)$
4. "P -1 security agent"(p-1 sa) requests Public key of EmP_id from "Customer Authentication (CA)" $Emp\_id_{k\_pub}$?
5. "CA" sent the Public key of EmP_id to "p-1 sa". $(Emp\_id_{k\_pub})_{p-1sa_{k\_pub}}$

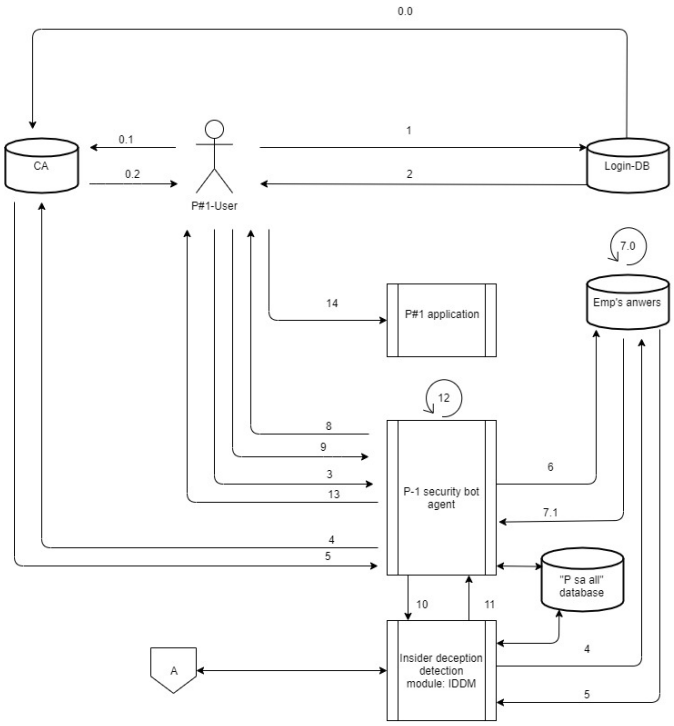

Fig. 1. ITDP sequence of tasks explanation

6. "p-1sa" sent EmP_id to "employee's stored answer database: ESA". *Emp _ id*
7. "employee's stored answer database" sent scrambled set of questions and answers ,{question$_i$} and {answer$_i$} : i=1,10, to "p -1 sa".
8. "p -1 sa" sent scrambled questions, {question$_i$}, under "p -1 application user" public key encryption. $(\text{Emp\_id}, \{\text{question}_i\}_{\text{Emp\_id}_{k\_pub}})_{p-1sa_{k\_priv}}$
9. "p -1 application user" answer all questions and fill his answers in the set of an answer, {answer´$_i$}, according to set of all {question$_i$}: i=1,10, to "p-1 sa". $(\text{Emp\_id}, \{\text{answer}_i'\}_{\text{Emp\_id}_{k\_priv}})_{p-1sa_{k\_pub}}$
10. "p -1 sa" sent {question$_i$}, {answer$_i$} and Emp_id 's{answer´$_i$} to "Insider Deception Detection Module: IDDM". $(\text{Emp\_id}, (\{\text{questioni}\}, \{\text{answeri}\}), \{\text{answer'i}\}, \text{start-time}, \text{stop-time})_{\text{IDDM}_{k\_pub}}$
11. "IDDM" process of "p -1 application user" 's {answer´$_i$} with {question$_i$} for authentication. "IDDM" sent deception scoring of "p-1 application user" back to "p -1 sa".
12. "p -1 sa" decides if "p -1 application user" should be permitted to get access to the P -1 application. The criteria of do not allow is depend on whether binary logistic regression of "Intruder" class variable score is greater than "0". The decision is made subject to "p -1 sa". Note, process 9th-12th might be iteratively performed not more than three times a trial.

13. If all answers are correct then "p -1 sa" sends a message "You are allowed to connect to the p-1 application".
    $((\text{"You are allowed to connect to p-1 application"},\text{stop-time})_{\text{p-1 application}_{\text{k\_priv}}})_{Emp\_id_{k\_pub}}$
14. Now, "p-1 application user" is allowed access to the p-1 application.

CA: Task Explanation

0.0 "Login DB" sends an encrypted message of "p-1 application user" under "IT-ad" attestation. $((\text{Emp\_ID})_{\text{Emp\_id}_{\text{K\_priv}}})_{\text{IT\_ad}_{\text{k\_priv}}}$

0.1 "p-1 application user" sends his public key to "CA". $(Emp\_id, K\_pub_{EMP\_ID})_{ca_{k\_pub}}$

0.2 "CA" recheck the message authentication attestation sent from "IT_ad", step #0.0. If the message $((\text{Emp\_ID})_{\text{Emp\_id}_{\text{K\_priv}}})_{\text{IT\_ad}_{\text{k\_priv}}}$ can be decrypt $Emp\_ID_{k\_pub}$ by revealing, Emp_ID then $(Emp\_id, K\_pub_{EMP\_ID})$ is kept in the CA database. Note, "Emp_id" is the same person that acts as "P-i application user" when he is assigned to "P -i application".

b) Insider Deception Detection Module: IDDM"

1. Emp_idi's data collection

1.1. Website logs data collection:

- IDDM: requests all emp_id's website connection history from the website logs database: WSL. The website logs data are composed of {Time, user name, URL of visited website}.
- All accumulated emp_id's website connection is prioritized to only the three most visiting websites based on the amount of access.
- $\{\text{Emp\_id}_i, \text{website\#1}, \text{website\#2}, \text{website\#3}\}$ is appended in the IDDM-WSL database. Note, the activity is periodically performed under IDDM's refreshing time policy.

1.2. Data processing logs collection

- IDDM: requests all of the emp_id's data processing from the data processing logs database. The processing logs are composed of {Emp_id$_i$'s, procedure name, start time, stop-time}.

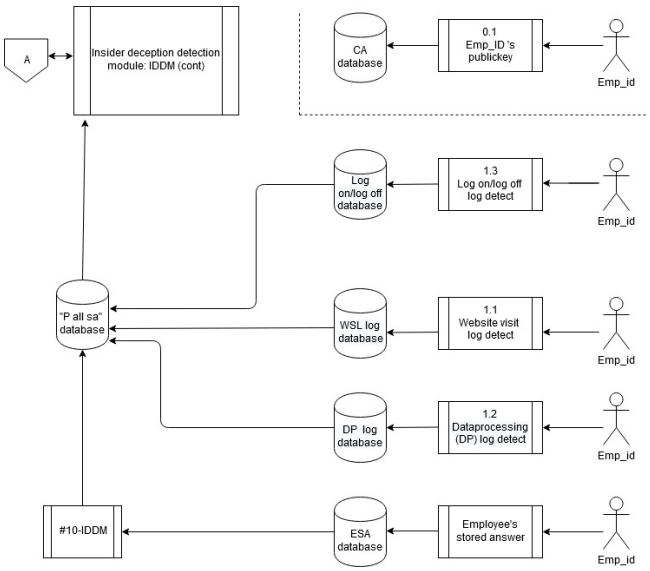

Fig. 2. IDDM sequence of tasks explanation

1.3. Log on-log off logs collection

- IDDM: requests all emp_id$_i$'s Log on-log off logs collection from computer log on logs database. The processing logs are composed of {Emp_id$_i$'s, log-on time, log off -time}.

1.4. Question-answering CI-time

- IDDM: ask all emp_id$_i$'s Question-answering time from the "p -all sa" database. The "p -all sa" database has its duty about keeping all emp_id$_i$'s Question-answering time measures. Whenever those emp_id$_i$'s request to access to some application – program #i, "p -all sa" perform IDDM#10(fig.1). "p-1 application user" completely answers all questions then set all answering back to "p_isa", IDDM#11(fig.1). Sending time and receiving time were kept in "p -all sa" database.

2. Data Record Preparation

2.1. Emp_id$_i$'s Web site access behavior

"Web access behavior" attribute is calculated on $\{Emp\_id_i, website\#1, website\#2, website\#3\}$ from "IDDM#1.1".

Data type of website #i is nominal such as "google.com", "youtube.com", etc. However, three frequently used web site should be altered according to emp_id$_i$'s website usage behavior. These calculated attributes are kept in the "p -all sa" database.

2.2. Emp_id$_i$'s Task processing average working CI-time.

"Task average working CI-time" attribute is calculated from IDDM#1.2; Since data processing time on each emp_id$_i$'s assigned application program (obligation) should take not an exact length of time to finish his task thus the average of data processing time is not suitable. History data processing time is transformed into a confidence time interval of data processing time. Task average working CI-time value is $\overline{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}; \alpha 0.05$. While, $\overline{X}$ is task average working time and σ is the standard deviation of task working time. These calculated attributes are kept in the "p -all sa" database. $\{\text{Emp\_id}_i, \text{Prog\#j}, \text{Task average working CI-time}\}$

2.3. Emp_id$_i$'s Working start & stop (log on & log off) confidence interval-time

Attributes "Working start CI-time" and "Working stop CI-time" are calculated from IDDM#1.3. These calculated attributes are kept in the "p -all sa" database. $\{\text{Emp\_id}_i, \text{Working start CI-time, Working stop CI-time}\}$

3. Data Preparation

3.1. Preparation of Insider Threat Detection Dataset

3.1.1 Sample observation

To create the first insider threat detection dataset, there are many activities to process.

a) Thirty application users were asked to choose their answers to 5 questions. Each question has 5 predefined static choices. The questions and their choice of answers are shown in table 1.

Table 1. Predefined question and choice of answer

	Sport	Music genre	National favorite food	Drinks	Social media
1	bowling	pop	vegetarian food	pure / table water	twitter
2	boxing	rock	noodle soup	carbonate water	facebook
3	football	classic	spicy shrimp soup	tea	line
4	tennis	hiphop	chicken in coconut soup	coffee	instagam
5	swimming	jazz	spicy green papaya salad	orange juice	pinterest

Each emp_id (30 persons) has to choose his favorite answer for every question (sport, music genre, national favorite food, drinks, and social media). For example, data record of emp_id$_1$, sport=boxing, music genre=jazz, national favorite food = noodle soup, drinks=orange juice, social media=line) is coded as {emp_id$_1$, 2, 5, 2, 5, 3}. Every data record for 30 persons was kept in Emp's answers database. Attributes "Question-answering CI-time" are calculated. These calculated attributes are kept in the "p -all sa" database $\{\text{Emp\_id}_i, \text{Question-answering CI-time}\}$.

b) Application program assigned to each emp_idi

Each emp_id_i has the responsibility of a particular application program. This obligation of everyone are kept in P_i-sa database: $\{emp\_id_i, proc_j\}$.

c) Observation preparation

c-1). This research was limited to study only three application programs.

The group of employees is obligated to a particular application program. Emp_id# (1-10) is assigned to be an IT user of the program#1 Emp_id#(11-20) is assigned to be an IT user of the program#2. Emp_id# (21-30) is assigned to be its use of program#3.

c-2). Every emp_id_i is asked to process his obligation application program about 30 times. This activity is performed to create and append their real behavior about task working start CI-time, task working stop CI-time and task average working CI-time to "data processing log database".

c-3). Every emp_id_i is asked to surf on his favorite webs.

This activity is performed to create and append their real behavior about Web accessing behavior to the "WSL database".

d) Security penetration test

Every emp_id is assigned by a researcher to intently attack others, not his obligation application program. Since everyone knows all questions and choices of an answer to each question, table 1, therefore they can guess the answer to each question, which was sent from the "P -isa". However, it is very difficult that "Emp_id" can choose the correct answer for each question. Since there are five sending questions from the "P-isa", the correct answering to all questions is about $(1/5)^5 = 0.00032$ or 0.032%. Therefore, he has to try out more times to correct answering on all "p#isa's questions", questions than authentication or real emp_id's processing. Since every "Emp_id" is an insider employee, their behavior is already collected as prior explained. However, each "Emp_id" rather has the same behavior. This distinction should be used to classify if he is an authentic "Emp_id" who responded to a particular application program. These assigned "Emp_id", who attacks not to his responsible application program, are called an insider threat. There are thirty observations of insider threats. The normal and attack activity observation is further used in insider threat classification model training.

3.1.2 Control and class attribute

Gathered data of each attribute are coding to an ordinal scale to be used in data model training and testing.

a). Correct question-answering CI-time

The "Correct question-answering CI-time" is transformed into three rating scales. For example, if emp_id_i's "Correct question-answering-time" is less than "-Correct question-answering CI-time" then conditional attribute "Question-answering CI-time" is set to "1". As an example, if "Correct question-answering-CI-time" of "Emp_id_i" is

3-6 minute. Suppose, some “Emp_id$_j$", fake "Emp_id$_j$" or "IT user", try to attack his non-obligation application, "p#k application”, if his “Correct question-answering-time” is greater than “Correct question-answering-CI-time” of real emp_id$_j$ then, such as ‘8’ minutes, current “Correct question-answering-CI-time” is set to “3”. If emp_id$_j$’s “Correct question-answering-CI-time” scale value are 1 or 2 then this emp_id$_j$ seems to be a real emp_id$_j$.

Table 2. “Correct question-answering-time” transformation

IF emp_idi’s “Correct question-answering-time”	.or.	THEN
-Correct question-answering CI-time”	+Correct question-answering CI-time”	“Correct question-answering CI-time” is set to
less than or equal	n/a	1
greater than	equal	2
n/a	greater	3

b). Working start CI-time

"Working start CI-time" is a conditional attribute that is used to decide if some IT user is logged on to the computer system as usual log on time. For example, if emp_id$_i$’s “Working start -time” is less than or equal to “-Working start CI-time” then conditional attribute “Working start CI-time” is set to “1”.

Table 3. “Working start -time” transformation

IF emp_idi’s “Working start -time”	.or.	THEN
-Working start CI-time	+Working start CI-time”	Working start CI-time” is set to
less than or equal	n/a	1
greater than	equal	2
n/a	greater	3

c). Working stop CI-time

"Working stop CI-time" is the conditional attribute that is used to decide if some "IT user" is log off from the computer system as usual log off time. For example, IT emp_id$_i$’s “Working stop -time” is less than or equal to “-Working stop CI-time” then conditional attribute “Working stop CI-time” is set to “1”.

Table 4. “Working stop CI-time” transformation

IF emp_idi’s “Working stop -time”	.or.	THEN
-Working stop CI-time	+Working stop CI-time”	Working stop CI-time” is set to
less than or equal	n/a	1
greater than	equal	2
n/a	greater	3

d). Task average working CI-time

"Task average working CI-time" is a conditional attribute that is used to decide if the length of processing time for his responsible task has as usual task processing time. For example, if emp_id$_i$’s “Task average working time” is less than or equal to "-Task average working CI-time" then the conditional attribute "Task average working CI-time" is set to "1".

Table 5. "Task average working CI-time" transformation

IF emp_idi's task average working time	.or.	THEN
-task average working CI-time	+task average working CI-time	task average working CI-time is set to
less than or equal	n/a	1
greater than	equal	2
n/a	greater	3

e). Web access behavior

Web access behavior conditional attribute is represented three "Emp_id$_i$'s" favorite website. Since every IT user might arbitrarily changes his behavior then trained data about web access behavior should not be the same as new Web access behavior which is detected by the WSL-log database. For example, WSL- log database of "Emp_id$_i$" is {emp_id$_i$, Google, Facebook, Line} while the current WSL-log is {emp_id$_i$, Pinterest, Facebook, BBC news}. From a prior example record "google" is the most favorite website, so that rank data is given as "3".

Since the data type of "web access behavior" is ordinal then its value could be transformed into a quantitative variable through the normalization technique. After that, many dissimilarity measurement techniques such as Euclidean distance, "Chebyshev" distance, etc. are chosen to calculate for two objects' dissimilarity.

Table 6. Original rank data of two object on WSL based on r (1, 2, 3, 4)

Object	google	line	pinterest	facebook	bbc
1	3	1	0	2	0
2	0	0	3	2	1

The rank data is transformed to standardized value (0 to 1) by $s = \frac{r-1}{R-1}$, While r=ordinal value and R=max value of "r". Based on table 6, r is 4 (0, 1, 2, 3) and R is max(r) or 4.

Table 7. Normalized rank data of two object on WSL

Observation	google	line	pinterest	facebook	bbc
1	4	2	1	3	1
2	1	1	4	3	2

Note, normalized rank data Google: object#1, $s = (4-1)/(4-1) = 1$. Likewise, Facebook#1 $s = (3-1)/(4-1) = 0.667$. Since Pinterest: object#1 and BBC: object#1 are not in three favorite visiting websites then their "s" value was set to "1".

Table 8. Normalized rank data of two object on WSL

Observation	google	line	pinterest	facebook	bbc
1	1	0.333333	0	0.666667	0
2	0	0	1	0.666667	0.333333

The Euclidean distance value of the two observations is “0.60”.

$$Nd_{1,2} = \frac{\sqrt{(1-0)^2 + (0.33-0)^2 + (0-1)^2 + (0.67-0.67)^2 + (0-0.33)^2}}{\sqrt{1^2 + 0.33^2 + 0^2 + 0.67^2 + 0^2} + \sqrt{0^2 + 0^2 + 1^2 + 0.67^2 + 0.33^2}} = 0.60 \quad (1)$$

While “Nd” is “Normalized Euclidean distance” of two objects is calculated from equation (2) as shown.

$$Nd_{o1,o2} = \frac{\sqrt{\sum_{1}^{n} (a_i - b_i)^2}}{\sqrt{\sum_{i=1}^{n} a_i^2} + \sqrt{\sum_{i=1}^{n} b_i^2}} \quad (2)$$

$Nd_{o1,o2}$ data has value between [0,1]. To simply calculation about IT user behavior thus $Nd_{o1,o2}$ data is organized into three groups. If "$Nd_{o1,o2}$” is less than or equal to “0.40” then “Web access behavior” =1.If “$Nd_{o1,o2}$” is greater than “0.40” or equal to “0.80” then “Web access behavior” =2. If “$Nd_{o1,o2}$” is greater than “0.80” then “Web access behavior” =3.

f). Intruder

Every emp_id who is assigned by a researcher to attack others not his obligated application program, is marked as an insider intruder class variable. In this research, thirty “Emp_id_i" was assigned to be a fake “Emp_id_j". Their mission was set to create an experimental security breach incident.

g). IDDM related attributes-dataset

All calculated attributes from 3.12, a), b), c), d), e), f) are kept in the "P all sa” database. Partial preparing and gathering data of all attributes are presented in table 9. Conditional and decision attribute with their data type is represented in table 10.

Table 9. Partial observation with their conditional and class attribute.

Observation#	Correct question-answering CI-time	Working start CI-time	Working stop CI-time	Task average working CI-time	Web access behavior	Intruder
1	1	1	2	1	1	n
2	2	1	2	1	1	n
3	2	1	2	1	2	n
4	1	1	2	1	1	n
…	…	…	…	…	…	…
59	3	3	2	3	3	y
60	3	3	3	3	3	y

Table 10. Conditional and class attributes of ITDP classification

Attribute	Data type	Data range	Note
Correct question-answering CI-time	ordinal	1,2,3	conditional
Working start CI-time	ordinal	1,2,3	conditional
Working stop CI-time	ordinal	1,2,3	conditional
Task average working CI-time	ordinal	1,2,3	conditional
Web access behavior	ordinal	1,2,3	conditional
Intruder	nominal	y, n	class

3.1.3. Confidence interval

Conditional attribute confidence interval calculation, α 0.05, are shown in table 11. This data scale or boundary is used to assign each attribute continuous data value to an ordinal type.

Table 11. Summary of CI of all attributes

Attribute	Xbar	SD	n	+CI	-CI	Scale	Range
Correct question-answering CI-time (CQAT)	7	2	30	7.72	6.28	CQAT<6.28	1
						6.28≤CQAT≤7.72	2
						CQA>7.72	3
Working start CI-time (WST)	30	10	30	33.58	26.42	WST<26.42	1
						26.42≤WST≤33.58	2
						WST >33.58	3
Working stop CI-time (WSpT)	15	5	30	16.79	13.21	WSpT<13.21	1
						13.21≤WSpT≤16.78	2
						WSpT>16.78	3
Task average working CI-time (TAWT)	45	12	30	49.29	42.14	TAWT<42.14	1
						42.14≤TAWT≤47.86	2
						TAWT >47.86	3

3.2. Model training phase:

ITDP data set has thirty records that represent a normal situation (real IT user: intruder=n). The other thirty records are assigned as an abnormal situation (fake IT user; intruder=y). The data model is tried out under the "ten folds" technique. Training observations and testing observation ratio is “80:20”.

3.2.1. Rough set classification

Five answers to five questions are set as a conditional attribute. A class variable is the "Application program", in which each "IT employee" is assigned as his responsibility (obligation). “Table 12" presents a partial answer for all questions that are kept in the "p -all sa" database. The Rough set technique is used to find out patterns of all authentic "IT users" selected answering in the "p -all sa” database (3.1.1). RSS, Rough set tool, presented that some set of an attribute is not important since it is not effective in pattern construction. Set of minimal attributes that are adequate in pattern generating

is called “reduct” set, {mg(music genre), nff (national favorite food), dr (drinks), sm (social media)}while SP(sport) is unnecessary attribute.

Table 12. Partial the "p -all sa” database about conditional and class attributes.

ID employee	Sport	Music genere	National favorite food	Drinks	social media	Obligation
1	1	4	3	2	2	p1
2	2	2	3	3	2	p1
3	2	2	1	3	1	p1
4	3	3	3	1	4	p1
5	3	1	2	1	2	p1
6	5	2	5	1	5	p1
7	4	2	4	1	5	p2
8	3	2	3	4	3	p2
9	1	1	1	1	5	p2
10	2	3	1	1	4	p2

Thirty IT user answering observations were used to generate lower approximation patterns, table 13. Rule # 1, 6, and 9 are pointed to more application programs thus this situation should cause possible vulnerability.

Since in rule #1 there is two "IT user” give the same answers thus there exist some “IT user” whoever could act like another one. If he knew the secret key of another one then he could log-in computer system and answer the questions with his own set of answers. Unfortunately, the "p -i sa" allow this to imitate IT user’s access to another one obligation application program.

Table 13. Lower approximate rule on class variable “obligation-program”

Number	Rule	Support
1	(mg=4)&(nff=3)&(dr=2)&(sm=2)=>(class={p1[1],p3[1]})	2
2	(mg=2)&(nff=3)&(dr=3)&(sm=2)=>(class=p1[1])	1
3	(mg=2)&(nff=1)&(dr=3)&(sm=1)=>(class=p1)	1
4	(mg=3)&(nff=3)&(dr=1)&(sm=4)=>(class=p1)	1
5	(mg=1)&(nff=2)&(dr=1)&(sm=2)=>(class=p1)	1
6	(mg=2)&(nff=5)&(dr=1)&(sm=5)=>(class={p1[1],p2[1]})	2
7	(mg=2)&(nff=4)&(dr=1)&(sm=5)=>(class=p2)	1
8	(mg=2)&(nff=3)&(dr=4)&(sm=3)=>(class=p2[1])	1
9	(mg=1)&(nff=1)&(dr=1)&(sm=5)=>(class={p2[1],p3[1]})	2
10	(mg=3)&(nff=1)&(dr=1)&(sm=4)=>(class=p2)	1
11	(mg=2)&(nff=3)&(dr=5)&(sm=2)=>(class=p2)	1
12	(mg=2)&(nff=3)&(dr=1)&(sm=2)=>(class=p2)	1
13	(mg=2)&(nff=2)&(dr=4)&(sm=3)=>(class=p2)	1
14	(mg=1)&(nff=5)&(dr=1)&(sm=5)=>(class=p2)	1
15	(mg=4)&(nff=2)&(dr=2)&(sm=2)=>(class=p3)	1
16	(mg=2)&(nff=2)&(dr=3)&(sm=2)=>(class=p3)	1
17	(mg=3)&(nff=2)&(dr=2)&(sm=5)=>(class=p3)	1
18	(mg=2)&(nff=2)&(dr=2)&(sm=1)=>(class=p3)	1
19	(mg=2)&(nff=2)&(dr=1)&(sm=2)=>(class=p3)	1
20	(mg=4)&(nff=3)&(dr=1)&(sm=2)=>(class=p3)	1
21	(mg=2)&(nff=3)&(dr=2)&(sm=3)=>(class=p3)	1
22	(mg=2)&(nff=3)&(dr=4)&(sm=1)=>(class=p3)	1
23	(mg=1)&(nff=2)&(dr=2)&(sm=1)=>(class=p3)	1
24	(mg=2)&(nff=3)&(dr=5)&(sm=5)=>(class=p3)	1
25	(mg=2)&(nff=2)&(dr=1)&(sm=1)=>(class=p3)	1
26	(mg=2)&(nff=3)&(dr=2)&(sm=2)=>(class=p3)	1
27	(mg=1)&(nff=2)&(dr=2)&(sm=4)=>(class=p3)	1

This research suggests a solution to overcome this weakness by more concern about the computer usage behavior of each “IT user”. Partial “arff" file composed of five "Question-answering" attributes five "computer usage" attributes and class variable (obligation) as shown.

```
@RELATION B_C_obli
@ATTRIBUTE sp {1,2,3,4,5}
@ATTRIBUTE mg {1,2,3,4,5}
@ATTRIBUTE nff {1,2,3,4,5}
@ATTRIBUTE dr {1,2,3,4,5}
@ATTRIBUTE sm {1,2,3,4,5}
@ATTRIBUTE Correct_question_answering {1,2,3}
@ATTRIBUTE Working_start_CI_time {1,2,3}
@ATTRIBUTE Working_stop_CI_time {1,2,3}
@ATTRIBUTE Task_average_working_CI_time {1,2,3}
@ATTRIBUTE Web_access_behaviour {1,2,3}
@ATTRIBUTE class {p1,p2,p3}
@DATA
```

```
1,4,3,2,2,1,1,2,1,1,p1
.
.
.
3,2,3,4,3,2,1,2,2,2,p2
```

Thirty authentic "IT users"' answers and "computer usage" sought for their pattern by Rough set technique. The twenty-nine lower approximation patterns are shown in table 14. There are no "IT users" perform a similar pattern. Therefore, some IT users could not take another obligation assuming his name. In brief, this research should use the attribute "Correct question-answering CI-time" attribute substitute for "IT user's answers" attribute since "Correct question-answering CI-time" is a final goal of the "IT user's answers" function.

Table 14. Twenty-nine lower approximate rules on "IT user's answers", "Computer usage" and class variable "obligation-program"

Number	Rule	support
1	(dr=3)&(Working_stop_CI_time=2)=>(class=p1[2])	2
2	(sp=3)&(Task_average_working_CI_time=1)=>(class=p1[2])	2
3	(sm=5)&(Correct_question_answering=2)&(Working_stop_CI_time=2)=>(class=p2[4])	4
4	(dr=1)&(Correct_question_answering=2)&(Working_stop_CI_time=2)=>(class=p2[4])	4
5	(dr=1)&(Correct_question_answering=2)&(Task_average_working_CI_time=1)=>(class=p2[4])	4
6	(Correct_question_answering=1)&(Task_average_working_CI_time=3)=>(class=p2[2])	2
7	(Working_stop_CI_time=2)&(Task_average_working_CI_time=2)=>(class=p2[2])	2
8	(Working_stop_CI_time=2)&(Task_average_working_CI_time=3)=>(class=p2[2])	2
9	(sm=2)&(Working_start_CI_time=2)=>(class=p2[2])	2
10	ct_question_answering=2)&(Working_start_CI_time=1)&(Task_average_working_CI_time=1)=	3
11	(sp=3)&(sm=3)=>(class=p2[2])	2
12	(Task_average_working_CI_time=2)&(Web_access_behaviour=2)=>(class=p2[2])	2
13	(Task_average_working_CI_time=3)&(Web_access_behaviour=1)=>(class=p2[2])	2
14	(Working_stop_CI_time=1)=>(class=p3[7])	7
15	(mg=2)&(nff=2)&(sm=2)=>(class=p3[2])	2
16	(dr=2)&(Correct_question_answering=2)=>(class=p3[5])	5
17	(nff=2)&(dr=2)=>(class=p3[5])	5
18	(mg=2)&(nff=2)&(dr=1)=>(class=p3[2])	2
19	(mg=2)&(dr=2)=>(class=p3[3])	3
20	(sp=2)&(Correct_question_answering=1)=>(class=p3[2])	2
21	(sp=4)&(Correct_question_answering=1)=>(class=p3[2])	2
22	(nff=2)&(Correct_question_answering=2)=>(class=p3[5])	5
23	(sm=1)&(Working_start_CI_time=2)=>(class=p3[2])	2
24	(sp=1)&(sm=2)&(Correct_question_answering=2)=>(class=p3[2])	2
25	(sp=4)&(Task_average_working_CI_time=1)=>(class=p3[2])	2
26	(sm=1)&(Web_access_behaviour=1)=>(class=p3[4])	4
27	(mg=2)&(nff=2)&(Task_average_working_CI_time=1)=>(class=p3[3])	3
28	(mg=4)&(Correct_question_answering=2)=>(class=p3[2])	2
29	(Task_average_working_CI_time=2)&(Web_access_behaviour=1)=>(class=p3[3])	3

3.2.2. Insider threat classification

Partial data about computer usage (3.1.2) of thirty observations of authentic "IT user" (not intruder) and thirty observations of imitate "IT users" are presented in table 15. This dataset was used to find out the best classifier on the decision tree and the Discriminant analysis technique.

Table 15. Partial observations of "IT user computer usage of both "Intruder" and "Not Intruder".

Conditional attribute						Class attribute
Observation#	question-answering	Working start CI-time	Working stop CI-time	Task average working CI-time	Web access behavior	Intruder
1	1	1	2	1	1	n
2	2	1	2	1	1	n
3	2	1	2	1	2	n
4	1	1	2	1	1	n
…	…	…	…	…	…	…
59	3	3	2	3	3	y
60	3	3	3	3	3	y

3.2.2.1. Decision tree j48

From training data (3.2.2), a decision tree is tried out with many classification - tree algorithms on data mining tools, and WEKA 3.6.9. The best classifier is j.48 with an accuracy of about 91.67%. While with Random Forest, ID3 accuracy of classification is 85.33% and 77.58% respectively.

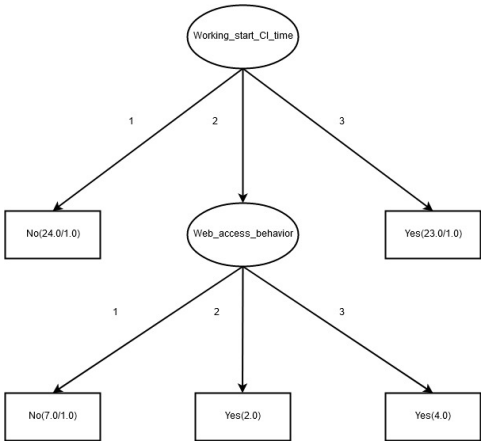

Fig. 3. J48 Decision tree

3.2.2.2. Linear binary discriminant function analysis

Linear binary Discriminant function analysis was calculated on five hundred iteration boots trap datasets. All Discriminant coefficients are significant at α0.05. There is a 98.3% correct classification. Function at group centroid of "not intruder" is "-2.768" and "+2.768 for "intruder" dependent variable. The cutting point of "not intruder" is considered if the Discriminant coefficient value is "≤0".

$$Intruder = -8.875 + 1.283cqa + 0.391taw + 0.924wab + 0.936wstart + 0.872wstop \quad (3)$$

4 ITDP Result and Evaluation

The result of rules construction from j48 decision tree classification (3.2.2.1) and binary Discriminant function analysis (3.2.2.2) gave a high accuracy in insider threat classification. There is an easy judgement if requesting “IT user” is an intruder by first considering on attribute "Wstart”. If its value is “3” then the guest is defined as an intruder since the "intruder” score of “Linear binary Discriminant function analysis” is less than “0” (-2.597) when “cqa", "taw", "wab” and “wstop” have value “1”.

On the other hand, binary Discriminant function analysis (3.2.2.2) is more preferably used by the "p -i sa” administrator. Since Discriminant function give a Discriminant score which “p-i security bot agent” could use it to consider the certainty of an intruder in continuous digit number while decision tree present certainty of intruder class variable in dichotomous nominal value (Yes or No).

5 Research Summary and Suggestion

ITDP is designed and tried out to detect and prevent insider threats. This protocol was evaluated by thirty IT users. The result of the evaluation found that ITDP could enhance capability on insider threat detection. ITDP could increase trustworthiness. Nevertheless, service performance is diminished. All IT users have to do checking on an assigned question-answering process. However, it is worthwhile especially on accessing to a sensitive organizational application.

6 Acknowledgement

This research was conducted at Information Technology Computer Science department of Chandrakasem Rajaphat University. “IT user” in this research, is chosen from fourth-year students and other client users. Their answers and website visits were kept in the various database and log databases. Log data were collected under the permission of the computer and network administrator. With all of their helping hands, the experiment was successfully undertaken.

7 References

[1] Z.Pawlak, Rough Sets - Theoretical Aspect of Reasoning about Data, Kluwer Academic Publishers, 1991.
[2] Huberty, C. J. and Olejnik, S. Applied MANOVA and Discriminant Analysis, Second Edition. Hoboken, New Jersey: John Wiley and Sons Inc., 2006.
[3] Shafi Gold Wasser and Mihir Bellare, Cryptography, University of California, July 2008.
[4] Hants Kipper, Strategies and techniques of questioning effectuating thinking and deep understanding in teaching engineering at Estonian Center for Engineering Pedagogy, University of Technology, Estonia, 2010.

[5] Frank L. Greitzer et al, SOFIT: Sociotechnical and Organizational Factors for Insider Threat, 39thIEEE Symposium on Security & Privacy, Workshop on Research for Insider Threat (WRIT), San Francisco, CA, 2018. https://doi.org/10.1109/spw.2018.00035
[6] Zakaria Suliman Zubi and Mussab Saleh El Raiani, Computer usage behavior using weblogs dataset via web mining for user behavior understanding, International journal of computers and communications, Vol. 8, 2014.
[7] Axelrad and Paul J. Sticha, A Bayesian Network Model for Predicting Insider Threats, Human resource research, USA, 2013. https://doi.org/10.1109/spw.2013.35
[8] Florian Kamm¨uller and Christian W. Probst, Modeling and Verification of Insider Threats using logical analysis, London, 2019.
[9] Judee Burgoon et al, An Approach for Intent Identification by Building on Deception Detection, Proceedings of the 38th Hawaii International Conference on System Sciences, USA, 2005.
[10] Wibisono Sukmo Wardhono et al, End-to-End Privacy Protection for Facebook Mobile Chat based on AES with Multi-Layered MD5, International Journal of Interactive Mobile Technologies, (iJIM), Vol. 12, No 1, 2018. https://doi.org/10.3991/ijim.v12i1.7472
[11] Tanweer Alam and Mohamed Benaida, Cloud–Internet Communication Security Framework for the Internet of Smart Devices, International Journal of Interactive Mobile Technologies (iJIM), Vol. 12(6).
[12] Pavani V L, A Novel Authentication Mechanism to Prevent Unauthorized Service Access for Mobile Device in Distributed Network, International Journal of Interactive Mobile Technologies (iJIM), Vol. 12, No. 8, 2018. https://doi.org/10.3991/ijim.v12i8.8194
[13] Ibrahim Obeidat, Intensive Pre-Processing of KDD Cup 99 for Network Intrusion Classification Using Machine Learning Techniques, International Journal of Interactive Mobile Technologies(iJIM), Vol. 13, No. 1, 2019. https://doi.org/10.3991/ijim.v13i01.9679
[14] Farida Achemoukh, Integration of User Profile in Search Process according to the Bayesian Approach, International Journal of Recent Contributions from Engineering, Science & IT (iJES), Vol. 6, No. 4, 2018. https://doi.org/10.3991/ijes.v6i4.9716

8 Authors

Amnat Sawatnatee, Ph.D. is a computer science lecturer at Chandrakasem Rajabhat University, Thailand. His research is about Multimedia and Data mining. amnat.s@chandra.ac.th

Somchai Prakancharoen, Ph.D. is a Faculty of Applied Science at King Mongkut's University of Technology, North Bangkok, Thailand. His interest research topics are computer security and Data mining.

Article submitted 2020-09-06. Resubmitted 2020-11-03. Final acceptance 2020-11-29. Final version published as submitted by the authors.