Analysis of Insider Access Pattern for Monitoring Misuse in the DCD

Jung ho Eom[1], Sung hwan Kim[2] and Tai Myoung Chung[2]

[1]Military Studies, Daejeon University, 62 Daehakro, Dong-Gu, Daejeon,
[2]Internet Management Technology Laboratory,
School of Information and Communication Engineering,
Sungkyunkwan University, Chunchun-dong 300,
eomhun@gmail.com, shkim47@imtl.skku.ac.kr, tmchung@ece.skku.ac.kr

Abstract

In this paper, we analyzed insider access pattern to documents for monitoring insider misuse in the document control domain (DCD). We optimized insider access pattern to documents using apriori algorithm adding time and click ratio. Insiders have typical access patterns among related documents when performing their duties. When insider misuse occurs in the DCD, it can be detected by the typical access pattern of insider. We added time and click ratio to apriori algorithm for improving the accuracy of insider access patterns. Time means the activation time of the document. Click ratio represents the frequency of click change within the document. The proposed algorithm can remove access patterns with the low reliability and reduce the false detection rate when detecting insider misuse.

Keywords: *Insider Misuse, Insider Threat, Insider Access Pattern*

1. Introduction

2011 CyberSecurity watch survey [1] said that 46% of respondents answered damage caused by insider attacks are more damaging than outsider attacks. And 63% of respondents answered that most common insider e-crimes are unauthorized access to data and use of corporate information. Damage by insider attack is greater than the outsider attack because insider knows well network and system information and the main location of the critical data. Also it is because they sometimes access the desired data with legitimate authorization with no doubt. The insider threat has emerged as one of the most important security concerns of business and government organizations. The insider threat is considered the most serious risk to system and network security. Insider threat is defined as an insider's behavior that puts at risk an organization's information, processes, or resources in a disruptive or unwelcome method [2-4].

In particular, data leakage by insiders has become an important issue in enterprise and government. For example, in U.S, it was hot issue that former Goldman Sachs Group director leaked corporate secrets related a hedge fund [5]. He has taken a personal financial gain as provided the leaked secret to competing company. Motivations of insider who leaks internal information are usually financial gain, business advantage, and revenge to his/her organization, and so on. In the past, data leakage protection techniques were mainly based on firewall, intrusion detection system, digital right management (DRM) and encryption [6]. But it is not effective to block the leakage of digital documents information by insider with legitimated access authorization through the old fashion security techniques. It includes privacy information, corporate confidential, the core technology and military secret in documents.

We proposed the analysis method of insider access to digital documents for optimizing access pattern. We optimized insider access pattern by improved apriori algorithm adding time and click ratio. Apriori algorithm is a method for frequent itemset mining and association rule over the data. We added time and click ratio to apriori algorithm for improving accuracy of access pattern.

In this paper we will describe related works in Section 2 and our proposed algorithm in Section 3. We explain application of proposed algorithm in Section 4 and conclude in the last section.

2. Related Works

2.1. Insider Threats

In the research report [7], Dawn Cappelli *et al.*, define malicious insider that is a current or former employee, contractor, or business partner who has or had authorized access to an organization's network, system, or data and intentionally exceeded or misused that access in a manner that negatively affected the confidentiality, integrity, or availability of the organization's information or information systems. Motivations for malicious behavior by insiders are as follows [7-9].

- IT sabotage: occurs when a disgruntled employee causes intentional damage to a database, systems, and business operations.

- Theft or modification for financial gain: secretly sell crucial proprietary data to an outsider with the hope of a monetary reward.

- Theft or modification for business advantage: steal confidential or proprietary information from the organization with the intent to use it for a new job or their own business.

Research report [7] indicated that 58% of the 176 insider incidents involved theft or modification of information. Insiders are non-technical and low level positions with access to confidential or sensitive information, programmers, and engineers, etc. They accessed customer information, intellectual property, and identifiable information with authorized access at their workplace during working hours. Insiders can bypass physical and technical security measures designed to prevent unauthorized access. Insiders can access the data with authorized access and same methods daily. Most security systems can detect attacks from outside threats, not necessarily internal threats. Insiders know the security policies, procedures, and the associated vulnerabilities. They can access to critical data as using security flaws in the system without doubt [8].

In recent, research on insider threat and security has been actively conducted. Research focuses on prevention, detection, monitoring, and forensics. Frank L. Greitzer *et al.*, [10] researched predictive modeling for insider threat mitigation. Their proposed predictive model focuses on a possible structure of a predictive model combining psychosocial and traditional digital data. A major challenge of this research is to define possible precursors to insider threat exploits in terms of observable cyber and psychosocial indicators and integrating these indicators in an analysis model. You Chen *et al.*, [11] proposed the community anomaly detection system for detecting anomalous insiders in collaborative information systems. This system is an unsupervised learning framework to detect insider threats based on the access logs. The framework is based on the insider access pattern. The proposed system consists of two components; relational pattern extraction and anomaly prediction. The former derives community structures.

The latter leverages a statistical model to determine a point time of deviation from communities. Aung H. Phyo *et al.*, [12] architected a framework for monitoring insider misuse of IT applications. The proposed framework combined the concept of RBAC to provide knowledge of role relationships with anomaly based detection techniques to effectively detect insiders who abuse their authorization. Recently, a document-based DRM is used for preventing the information leakage of electric documents by insiders. DRM has been developed for copyright security and piracy prevention of digital contents. This has been used as a means to prevent illegal access to a document or block internal leakage of the document [3]. In this paper, we focus on an analysis of insider access pattern optimization for using detection technique of insider misuse.

2.2. Apriori Algorithm

Apriori algorithm is an influential algorithm for mining frequent itemsets for association rules. Apriori algorithm finds frequent itemsets according to a user-defined minimum support. Frequent itemsets can be used to determine association rules which normalize typical patterns in the database [13]. In the first step in the algorithm, it constructs the candidate 1-itemsets. And then it generates the frequent 1-itemsets by deleting some candidate 1-itemsets are lower than the minimum support. Support means the probability of two items at the same time [13, 14].

$$Support(A \Rightarrow B) = P(A \cap B) \quad (1)$$

After it finds all the frequent l-itemsets, it joins the frequent 1-itemsets with each other to construct the candidate 2-itemsets. As this step is repeated until candidate itemsets can't be created, we can finally the last frequent itemsets. Support has the disadvantage could not properly measured the association when support value is small. In order to supplement the support's disadvantage, confidence is applied. Confidence means the occurrence probability of B after A.

$$Confidence(A \Rightarrow B) = P(B/A) = P(A \cap B)/P(A) \quad (2)$$

In order to construct all association rules, support is firstly calculated in the itemsets, and then confidence is calculated.

Apriori algorithm isn't included the frequent itemsets which relatively less occurred in the frequent itemsets for searching association rules. But it should be regarded as the frequent itemsets if usage time is long, even if item occurred relatively less than other, because time has more association in itemsets rather than frequency.

3. The Proposed Algorithm

3.1. Time and Click ratio based Apriori Algorithm

Our proposed algorithm can optimize insider access pattern using time and click ratio based Apriori algorithm. We added time factor to apriori algorithm. Time weight cited 'time weight per unit' in Kim's paper [15] as follow table.

Table 1. Time Weight

Unit	Time(Min)	Weight	Unit	Time(Min)	Weight
1	0 ~ 0.9	0	3	11 ~ 30	0.3
2	1 ~ 10	0.2	4	31 ~	0.5

Let supposed to do not use the document after insider activates the document. It also should consider including this item to frequent itemsets. If insider does not use the document for a long time, it acts as a negative factor to analyze insider access pattern. It is not easy to determine whether the documents use or not. In this paper, we introduce click ratio for checking whether the documents use or not. Click ratio defines as the frequency of click on the activated document [16]. But we limited click ratio to the changed frequency of selected text per page. In other words, it is that how many times changed during the activation time. The type of click includes all count by mouse and keyboard. Click ratio is calculated as follows.

$$Click\ Ratio = Click\ Count / Execution\ Time \quad (3)$$

We refer to the user log file for usage time and click ratio. Whenever insider activates the document in the DCD, log server saves all usage history which includes the type of activated document, open/close time, and click count, *etc*.

If time and click ratio applies to apriori algorithm, the time and click ratio based apriori algorithm (TCA) is as following formula.

$$TCA\ Value = (T+C) * P(A \cap B) \quad (4)$$

If there are many items in a itemset, TCA formula is as following.

$$TCA\ Value = [(\sum T_i / n) + (\sum C_i / n)] * P(A \cap B) \quad (5)$$

3.2. Application of Proposed Algorithm

Figure 1 shows an example of the movement path of document access by insiders. Insiders accessed to total 10 documents.

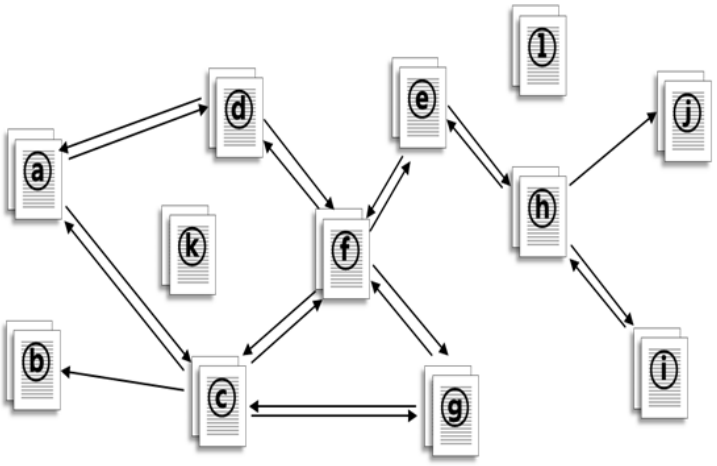

Figure 1. The Access Route by insider among Documents

In the Figure 1, the one-way arrows mean the closed document after working, and the document does not have arrows means that it does not access like ‘l’. The usage time of the document ‘b’ is 20 seconds, so it is deleted. Table 2 shows the candidate itemsets deleted documents ‘b, l’.

Table 2. Candidate Itemsets

Item Number	Candidate Itemsets
2	{a, c}, {c, d}, {c, f}, {c, g}, {d, a}, {d, c}, {d, f}, {f, c}, {f, d}, {f, e}, {f, i}, {e, f}, {e, h}, {g, f}, {h, e}, {h, i}, {i, f}
3	{a, c, d}, {c, d, f}, {c, f, d}, {c, g, f}, {d, a, c}, {d, c, f}, {d, f, c}, {f, c, d}, {f, d, c}, {f, c, g}, {g, f, c}
4	{a, c, f, d}, {c, g, f, d}, {c, f, d, a}, {c, d, f, d}, {d, a, c, f}, {d, f, e, f}, {d, c, g, f}, {f, d, a, c}, {f, d, c, g}, {f, e, h, i}, {e, h, i, f}, {h, i, f, e}, {i, f, e, h}
5	{a, c, g, f, d}, {c, g, f, d, a}, {d, a, c, g, f}

Table 2 shows total 52 candidate itemsets among documents in the DCD. For example, let supposed to there are activated documents and movement path by insider 'A' among documents in the DCD. If then, we can filter only insider 'A' related candidate itemsets in candidate itemsets

Table 3. Frequent Itemsets for Insider A

Item Number	Frequent Itemsets
2	{a, c}, {c, f}, {c, g}, {d, a}, {f, e}, {f, i}, {e, f}, {g, f}, {i, f}
3	{a, c, d}, {c, f, d}, {c, g, f}, {d, a, c}, {f, c, d}, {f, d, c}, {f, c, g},
4	{a, c, f, d}, {c, g, f, d}, {d, a, c, f}, {d, c, g, f}, {f, d, c, g}, {f, e, h, i}, {e, h, i, f}, {i, f, e, h}
5	{a, c, g, f, d}, {c, g, f, d, a}, {d, a, c, g, f}, {f, d, a, c, g}

Log server offers the history of access to documents by insider A. We arrange frequent intemsets including time and click count as follows Table 4.

TCA values calculated using equation 4. If we calculate TCA value of itemset {a, c}, all frequent itemsets including itemset {a, c} should be sum up. Itemset {a, c}'s frequent itemsets are [{a, c}, {a, c, d}, {a, c, f, d}, {a, c, g, f, d}]. In other word, {a, c}'s TCA value is as followings.

Time value = (0.3+0.3+0.3+0.2+0.5+0.3+0.5+0.3)/8 = 0.34

Click Ratio value = (0.25+0.67+0.17+0.1+0.1+0.1+0.1+0.1+0.1)/8 = 0.21

{a, c}'s TCA value (0.34+0.21)*4/5=0.44.

Table 5 shows the calculation result of item number 2 in the frequent itemsets. If we process this calculation procedure, we can get the result until item number 5. In this paper, we limited to itemset 2. If minimum threshold is 60%, itemsets({a,c}, {c,f}, {f,e}, {e,f}) is dropped.

Table 4. Frequent Itemsets for Insider A

Item Number	Frequent Itemsets {A(time, click count)}
2	{a(20,5), c(30,20)}, {c(30,15), f(30,20}, {c(25,10), g(15,10)}, {d(30,15), a(10,5)}, {f(25,10), e(10,5)}, {f(50,30), i(30,20)}, {e(40,20), f(50,25)}, {g(60,30), f(20,10)}, {i(60,40), f(10, 10)}
3	{a(30,5), c(10,1), d(10,2)}, {c(60,20), f(30,15), d(20,10)}, {c(20,2), g(10,5), f(30,5)}, {d(20,5), a(15,10), c(20,5)}, {f(50,20), d(30,20), c(10,4)}, {f(60,6), c(30,6), g(20,5},
4	{a(40,4), c(20,2), f(20,5), d(30,5}, {c(40,10), g(10,5), f(30,5), d(20,10)}, {d(30,5), a(20,5), c(20,5), f(20,5)}, {d(30, 10), c(20,2), g(10,5), f(30,5)}, {f(30,15), d(15,5), c(10,4), g(10,5)}, {f(20,10), e(10,5), h(10, 1), i(15, 1)}, {h(20, 2), i(15, 1), f(25,10), e(10,5}, {i(60,40), f(10, 5), e(10,5), h(10, 2)}
5	{a(40,4), c(20,2), g(20,10), f(30,5), d(20,10)}, {c(30,10), g(20,4), f(30,5), d(20,5), a(10,5)}, {f(50,20), d(15,5), a(20,5), c(30,5), g(20,5)}

Table 5. The Results of TCA Value

Itemset	TCA Value	Itemset	TCA Value	Itemset	TCA Value
{a,c}	0.43	{d,a}	0.65	{e,f}	0.25
{c,f}	0.55	{f,e}	0.58	{g,f}	0.78
{c,g}	1.08	{f,i}	1.04	{i,f}	1.06

The results of TCA value compared with Apriori and time based Apriori is shown in the following Table 6. The formula of time based Apriori algorithm is as followings.

$$\textit{Time based Apriori algorithm value} = (\sum Ti) * P(A \cap B) \quad (\mathbf{6})$$

We know that the results of time based Apriori algorithm usually decided by usage time of itemset. TCA value is smaller than value by time based apriori algorithm. It is demonstrated that click ratio plays important role to optimize insider access pattern among documents. Click ratio indicates how often an insider changed selected text by cursor in the activated document. It means that usage of the document is high. If minimum threshold is decided 0.60, 2 itemsets are dropped by apriori algorithm. In the time-based apriori algorithm, 1 itemset is dropped. In our proposed algorithm, 4 itemsets are dropped. The threshold can be adjusted for the accuracy of the access pattern at any time. We can't confirm to improve the accuracy of the access pattern as reducing the number of itemset. But our proposed algorithm can be formalized insider access pattern than apriori algorithm. The proposed algorithm can reduce the low reliability and the false detection rate than apriori algorithm and time based apriori algorithm when detecting insider misuse.

Table 6. The Comparison of Apriori, Time-based and TCA

Itemset	Apriori	Time-based Apriori	TCA Value
{a,c}	0.8	2.16	0.43
{c,f}	0.8	2.08	0.55
{c,g}	0.8	1.52	1.08
{d,a}	1	2.8	0.65
{f,e}	0.8	1.52	0.58
{f,i}	0.2	1.6	1.04
{e,f}	0.2	0.2	0.25
{g,f}	1.2	3.96	0.78
{i,f}	0.6	1.2	1.06

Figure 2 shows the comparison of the results from apriori algorithm, time-based apriori algorithm and time and click ratio based apriori algorithm. When calculating until itemset number 5, the result will be more correct. If time and click ratio based apriori algorithm is used for detecting insider misuse, it will reduce the false detection rate. If the detection rate is low, minimum threshold could be controlled.

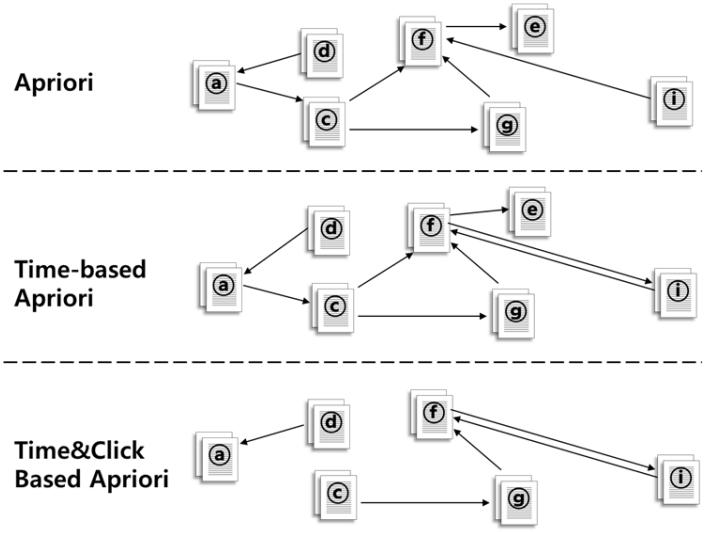

Figure 2. The Pattern comparision of Apriori, Time-based, and Time&Click based Algorithm

4. Conclusion

We proposed time and click ratio based apriori algorithm for analyzed insider access pattern to documents for monitoring insider misuse in the document control domain. Insiders have their typical access patterns to the documents when performing their duties. We optimized the movement path from document to document by insider according to the proposed algorithm. We added time and click ratio to apriori algorithm

for improving the accuracy of insider access patterns. We considered the usage time of document by time factor and the insider's behaviors on documents by click ratio. Time and click ratio is applied as weight factors to apriori algorithm. Click ratio plays important role to characterize access pattern to document. If only the activation time apply to the weights, support of a long inactivity time in the document will increase. So, we applied click ratio to time based apriori algorithm for complementing this disadvantage. Click ratio indicates how often an insider changes selected text in the activated document. This means that usage of the document is high. The proposed algorithm can reduce the low reliability and the false detection rate when detecting insider misuse.

In future, we continue to research the methodology of matching the progressing movement among documents to insider access pattern in the database. And we will apply the proposed algorithm to threat detection system for insider's misuse.

References

[1] 2011 Cyber Security Watch Survey, CSO magazine, U.S. Secret Service and Carnegie Mellon University&Deloitte, **(2011)** January.
[2] S. L. Pfleeger, J. B. Predd, J. Hunker and C. Bulford, "Insiders Behaving Badly: Addressing Bad Actors and Their Actions", IEEE Transactions on Information forensics and Security, vol. 5, no. 1, **(2010)**, pp. 169-179.
[3] J. -h. Eom, N. -u. Kim, S. -h. Kim and T. -m. Chung, "An Architecture of Document Control System for Blocking Information Leakage in Military Information System", International Journal of Security and Its Applications (IJSIA), vol. 6, no. 2, **(2012)** April, pp. 109-114.
[4] J. Hong, J. Kim and J. Cho, "The Trend of the Security Research for the Insider Cyber Threat", International Journal of Security and Its Applications, (IJSIA), vol. 4, no. 3, **(2010)** July, pp. 55-64.
[5] http://online.wsj.com/article/SB10001424052970203897404578077050403577468.html.
[6] J. J. Wu, J. Zhou, Jun Ma, S. Z. Mei and J. C. Ren, "An Active Data Leakage Prevention Model for Insider Threat", International Symposium on Intelligence Information Processing and Trusted Computing, IEEE press, **(2011)**, pp. 39-42.
[7] D. Cappelli, A. Moore, R. Trzeckak and T. J. Shimeall, "Common Sense Guide to Prevention and Detection of Insider Threats", Carnegie Mellon Software Engineering Institute, **(2009)**.
[8] J. White and B. Panda, "Automatic Identification of Critical Data Items in a Database to Mitigate the Effects of Malicious Insiders", ICISS 2009, LNCS, vol. 5905, **(2009)**, pp. 208-221.
[9] J. Kim, J. Hwang and H. -J. Kim, "Privacy Level Indicating Data Leakage Prevention System", International Journal of Security and Its Applications(IJSIA), vol. 6, no. 3, **(2012)** July, pp. 91-96.
[10] F. L. Greitzer, P. R. Paulson, L. J. Kangas, L. R. Franklin, T. W. Edgar and D. A. Frincke, "Predictive Modeling for Insider Threat Mitigation", Pacific Northwest National Laboratory Report, U.S Department of Energy, **(2009)**.
[11] Y. Chen, S. Nyemba and B. Malin, "Detecting Anomalous Insiders in Collaborative Information Systems", IEEE Transactions on Dependable and Secure Computing, vol. 9, no. 3, **(2012)**, pp. 332-344.
[12] A. H. Phyo, S. Furnell and E. Ifeachor, "A Framework for Monitoring Insider Misuse of IT Applications", Proceedings of Information Security South Africa 2004, **(2004)**, pp. 231-246.
[13] H. Kang, K. Yang, C. Kim, W. Rhee and B. Lee, "A Time-based Apriori Algorithm", Transaction of KIEE, vol. 59, no. 7, **(2010)**, pp. 1327-1331.
[14] Y. Ye and C. -C. Chiang, "A Parallel Apriori Algorithm for Frequent Itemsets Mining", Proceedings of the Fourth International Conference on Software Engineering Research Management and Applications (SERA'06), **(2006)**, pp. 87-93.
[15] J. Kim and J. Kim, "Analysis of Web User Access Pattern using Time based Association Rules", Proceedings of Korea Industrial Engineering Spring Conference, **(2001)**, pp. 799-802.
[16] S. -H. Kim, Y. -h. Choi, J. -h. Eom and T. -M. Chung, "Optimization of Insider Behavior Pattern for Detecting Misuse in the DCD", The proceedings of ISA 2013, (**2013**)

Authors

Jung ho Eom received his M.S. and Ph.D. degrees in Computer Engineering from Sungkyunkwan University, Suwon, Korea in 2003 and 2008, respectively. He is currently a professor of Military Studies at Daejeon University, Daejeon, Korea. His research interests are information security, cyber warfare, network security.

Sung hwan Kim received the M.S degree in Computer Science Engineering from Seoul National University, Seoul, Korea, in 2006. He is currently working toward the Ph.D. degree in the School of Information & Communication Engineering, Sungkyunkwan University, Suwon, Korea. His research interests are Information Security, Cyber warfare and SCADA Security.

Tai myoung Chung received his first B.S. degree in Electrical Engineering from Yonsei University, Korea in 1981 and his second B.S. degree in Computer Science from University of Illinois, Chicago, USA in 1984. He received his M.S. degree in Computer Engineering from University of Illinois 1987 and his Ph.D. degree in Computer Engineering from Purdue University, W. Lafayette, USA in 1995. He is currently a professor of Information and Communications Engineering at Sungkyunkwan University, Suwon, Korea. He is now a vice-chair of the Working Party on IS & Privacy OECD.