리력자료분석을 위한 새로운 빈발모임발굴방법

황석철, 윤룡한

자료에 대한 접근차단과 접근권한의 제한, 화일의 암호화/복호화를 진행하여 해당 단위의 내부비밀자료를 보호하는데 화일체계구동기(File System Driver)기술을 많이 리용하고 있으며 이 기술을 리용하는 과정에 얻게 되는 화일접근리력을 분석하여 유용한 정보를 추출하는 기술은 자료발굴분야의 중요한 연구분야이다.

선행연구[2]에서는 오림판(Clipboard)과 등록부변경리력을 기록하고 USB기억기로부터 복사하여 작업을 진행하던 자료들에 대한 삭제정형을 리력자료를 분석하여 확인하기 위한 방법에 대하여 연구하였지만 이 방법은 단순한 리력자료의 대조에 기초하고있다.

선행연구[4]에서는 기업체의 생산리력자료모임을 몇개의 속성들로 이루어진 업무자료기지로 변환하고 한번의 업무자료기지에 대한 순환탐색으로 빈발모임을 발굴하기 위한 방법을, 선행연구[3]에서는 관계형자료기지들과 홈페지접속에 대한 리력을 분석하는 방법을, 선행연구[1]에서는 대용량업무자료기지에서 련관규칙을 발굴하기 위한 빈발모임발굴 알고리듬을 제기하였다. 그러나 화일접근리력은 사용자의 간단한 조작에 대하여서도 많은 개수의 화일조작요청이 발생하며 하나의 화일조작에 따르는 리력은 실행되는 시점과 프로그람에 따라 기록되는 리력의 순서와 개수가 일치하지 않는 특성을 가지는것으로 하여 우와 같은 방식의 빈발모임발굴방법을 적용할수 없다.

론문에서는 화일에 대한 실행프로그람의 화일접근호출리력을 분석하여 사용자의 화일조작형태를 식별해낼수 있는 한가지 빈발모임발굴방법을 제기하였다.

화일열람기를 통하여 1.txt화일에 대한 복사조작(Ctrl+c)을 진행하였을 때 Explorer.exe의 화일접근리력의 일부는 그림과 같다.

0	0.0000142	1/17/2020 10:45:32 AM	CreateFile	C:\1.txt
1	0.0000037	1/17/2020 10:45:32 AM	QueryBasicInformationFile	C:\1.txt
2	0.0000063	1/17/2020 10:45:32 AM	CloseFile	C:\1.txt
3	0.0000268	1/17/2020 10:45:32 AM	QueryDirectory	C:\1.txt
4	0.0000149	1/17/2020 10:45:32 AM	CreateFile	C:\1.txt
5	0.0000027	1/17/2020 10:45:32 AM	QuerySecurityFile	C:\1.txt
6	0.0000017	1/17/2020 10:45:32 AM	QuerySecurityFile	C:\1.txt
7	0.0000067	1/17/2020 10:45:32 AM	CloseFile	C:\1.txt
8	0.0000122	1/17/2020 10:45:32 AM	CreateFile	C:\1.txt
9	0.0000026	1/17/2020 10:45:32 AM	QueryBasicInformationFile	C:\1.txt
10	0.0000057	1/17/2020 10:45:32 AM	CloseFile	C:\1.txt
11	0.0000123	1/17/2020 10:45:32 AM	CreateFile	C:\1.txt
12	0.0000026	1/17/2020 10:45:32 AM	QueryBasicInformationFile	C:\1.txt
13	0.0000057	1/17/2020 10:45:32 AM	CloseFile	C:\1.txt
14	0.0000146	1/17/2020 10:45:32 AM	CreateFile	C:\1.txt:Zone.Id...
15	0.0000149	1/17/2020 10:45:32 AM	QueryDirectory	C:\1.txt
16	0.0000192	1/17/2020 10:45:32 AM	QueryDirectory	C:\1.txt
17	0.0000146	1/17/2020 10:45:32 AM	QueryDirectory	C:\1.txt

그림. Explorer.exe의 화일접근리력

이 리력은 홈페지접속리력이나 자료기지접속리력과는 다른 특성을 가진다.

그림과 같이 화일을 실행하는 사용자의 간단한 조작에 대하여서도 ReadFile, WriteFile,

CreateFile, CloseFile과 같은 화일조작요청이 많이 발생하며 그에 따라 많은 리력이 기록되게 된다. 또한 같은 화일조작에 대하여 실행프로그람에 따라, 실행될 때마다 리력이 남는 형식과 순서가 일치하지 않는다.

우리는 이러한 리력자료로부터 사용자의 화일조작에 따르는 반복되는 패턴을 발굴하여야 한다.

정의 1 화일들에 대한 실행프로그람의 접근에 의하여 발생하는 매개 리력을 화일접근사건이라고 하고 화일접근사건들의 모임을 $E=\{e_0, e_1, e_2, \cdots, e_n\}$으로 표시한다.

그림 1과 같이 화일접근사건들은 여러개의 속성으로 이루어져있으며 발생한 순서에 따라 그 번호와 발생시간이 증가한다.

i째 사건을 번호와 시간까지 고려하여 $e_i=(en_i,\ sr_i,\ ts_i)$로 표시하자. 여기서 en은 화일접근요청, sr는 사건의 번호, ts는 사건이 발생한 시간이다.

실례로 그림과 같은 리력에서 $e_{11}=$(CreateFile, 11, 2020.1.17 10:45:32)이다.

정의 2 화일접근요청들로 이루어진 렬을 화일접근요청렬이라고 한다.

실례로 {CreateFile, ReadFile, CreateFile, CloseFile}은 길이가 4인 화일접근요청렬이다.

정의 3 E를 화일접근사건모임, α를 화일접근요청렬이라고 하자.

이때 α가 $\{en_0, en_1, \cdots, en_n\}$의 부분렬이면 α는 E에서 발생한다고 말한다.

정의 4 화일접근사건모임을 E가 $E_1, \cdots, E_k$로 분할되였을 때 화일접근요청렬 α가 발생하는 $E_i\ (1\le i\le k)$의 개수가 주어진 어떤 수보다 크면 α는 E의 분할 $E_1, \cdots, E_k$에서 빈발이라고 하며 그 길이가 i이면 i-빈발화일접근요청렬이라고 한다.

정의 5 두 화일접근렬에 대하여 적당한 자연수 k가 있어서 한 화일접근렬의 뒤로부터 k개와 다른 화일접근렬의 앞으로부터 k개의 화일접근요청들이 같으면 두 화일접근렬을 련결가능하다고 말한다.

다음으로 화일접근리력에서 빈발인 화일접근렬발굴방법에 대하여 보자.

걸음 1 화일의 복사, 이동, 이름변경, 삭제, 화일내용의 반출을 비롯한 사용자의 화일조작행위를 식별하기 위하여서는 우선 리력자료를 수집해야 한다. 이를 위하여 판별을 목적으로 하는 화일조작을 반복진행하여 리력자료를 수집한다.

걸음 2 얻어진 리력자료를 분할한다.

일반적으로 화일에 대한 접근리력은 0.000 1s사이에 발생하므로 두 리력이 기록된 시간차가 1s보다 크면 서로 다른 화일조작에 의하여 발생한 사건이라고 보아도 타당하다.

다음의 알고리듬을 리용하여 리력자료를 순환하면서 리력자료를 분할한다.

$k=0$

$E[0]=\{e_0\}$

$t=e_k$.datetime:

for ($i=1$; $i<n$; $i++$)

{

if (e_i.datetime $-t>1$)

{

k ++;

$i=e_i$.datetime;

}

else

{
$E[k] = E[k] \cup e_i$
}
}

걸음 3 빈발인 화일접근요청렬들을 구하기 위하여 얻어진 E의 분할 $E_1, \cdots, E_k$와 최소지지수 min_sup를 입력한다.

걸음 4 자료기지를 순환하면서 1－빈발화일접근요청렬들의 모임들을 구성한다.

걸음 5 $(i-1)$－빈발화일접근요청렬들의 모임족 F_{i-1}로부터 i－빈발화일접근요청렬들의 모임족 F_i를 발굴하기 위하여 F_{i-1}에 포함되는 두 $(i-1)$－빈발화일접근요청렬들의 가능한 모든 쌍 A, B를 선택한 다음 A, B의 가능한 모든 련결들중에 $E_1, \cdots, E_k$에서 빈발이고 길이가 i인 련결들을 F_i에 추가한다.

걸음 6 F_{i-1}로부터 F_i를 구성하는 과정을 F_i가 빈모임일 때까지 반복한다.

걸음 7 빈발인 화일접근요청렬들의 모임 $F = \bigcup_{j=1}^{i-1} F_j$를 출력한다.

표. 성능평가결과

리력 수집회수/번	검사 회수/번	정확한 판별회수/번	정확도/%
10	20	15	75.00
25	25	19	76.00
35	30	24	80.00
45	40	37	92.50
50	45	42	93.33

걸음 8 빈발인 화일접근요청렬들을 가지고 사용자의 화일조작을 판별한다.

다음으로 우리는 Windows화일열람기의 화일접근리력을 분석하여 Ctrl+c지령을 식별하는 방법으로 알고리듬의 성능을 평가하였다.

화일들에 대한 Ctrl+c지령을 반복진행하여 리력을 수집하고 최소지지도를 90%로 놓고 빈발모임을 탐색하고 그 패턴을 가지고 Ctrl+c조작을 식별해본 결과는 표와 같다. 표에서 보는바와 같이 리력수집회수가 증가할수록 해당한 화일조작에 대한 식별정확도가 높아진다는것을 알수 있다.

참 고 문 헌

[1] **김일성**종합대학학보 수학, 65, 2, 17, 주체108(2019).

[2] C. Ishizawa et al.; Int. J. Soc. Mat. Eng. for Res., 19, 11, 2013.

[3] Bo Li et al.; Journal of Systems Architecture, 81, 92, 2017.

[4] Y. Djenouri et al.; Knowledge-Based Systems, 139, 132, 2018.

주체109(2020)년 3월 15일 원고접수

A Novel Frequent Itemset Mining Approach for Log Data Analysis

Hwang Sok Chol, Yun Ryong Han

We propose a new method for mining frequent patterns from file access logs and apply it to identification of user's file action type. The proposed method can be applied to the other cases such as web log analysis and process log analysis.

Keywords: frequent itemset, file log analysis, file access log