군집화를 이용한 기업 핵심기술 유출자 분류에 관한 연구*

허승표** · 이대성** · 김귀남**

요 약

최근 국내 기업의 핵심기술 유출은 해마다 증가하고 있고 국가적인 차원에서 피해액 손실도 매년 크게 증가하고 있다. 최근 5년 간 우리나라 주요 사업의 기술유출에 따른 피해액은 220조 원에 달했고 2010년 총 예산과 비슷한 액수이다. 또한 핵심기술 유출 유형별로는 전직 직원, 현직 직원, 협력 업체 직원, 유치과학자, 투자업체 순으로 내부 인력에 의한 핵심기밀 유출이 가장 많은 것으로 나타났다. 이처럼, 사람에 의해서 핵심기밀 유출이 가장 많이 발생한 것에 따라 기업의 인원 보안 관리에 대한 대책 및 관리가 제대로 이루어 지지 않는 것을 미루어 짐작할 수 있다. 따라서 본 논문은 기업에서의 핵심기술 유출을 방지하기 위하여 내부 인력에 대한 사전 정보를 데이터 마이닝 방법을 통해서 핵심기술 유출 징후 분류 방법을 제안한다.

A Study on The Leak of Core Business Technologies Using Preventative Security Methods Such as Clustering

Seung Pyo Huh** · Dae Sung Lee** · Kui Nam Kim**

ABSTRACT

Recently, the leak of domestic core technology of major business in Korea and the subsequent damage, has been increasing every year. Financial losses due to this leak are estimated to be about 220 trillion, which is equivalent to the gross budget of Korea Besides, the majority of the leaks are caused by former and current staff members, cooperated businesses, scientists and invesment companies. This shows that the source of the leaks are internal personnel. In this manner, we can infer that the management and plan of personnel security has not implemented sound practices to prevent technology leak by people. Therefore, this thesis suggests classifying methods of technology leak through clustering, one of the data mining methods about the information of internal personnel to prevent core technology leak from businesses.

Key words : Industrial Security, Personnel Security, Clustering

접수일 : 2010년 7월 14일; 채택일 : 2010년 9월 15일

* 본 연구는 경기도 기술개발사업의 사업비지원(과제번호#A10101110)에 의해 수행되었음.

** 경기대학교 산업보안학과

1. 서 론

최근 국내 기업의 핵심기술 유출은 해마다 증가하고 있고 국가적인 차원에서 피해액 손실도 매년 크게 증가하고 있다. 최근 5년 간 우리나라 주요 사업의 기술유출에 따른 피해액은 220조 원에 달했고 2010년 총 예산과 비슷한 액수이다. 또한 핵심기술 유출 유형별로는 전직 직원, 현직 직원, 협력 업체, 유치과학자, 투자업체 순으로 내부 인력에 의한 핵심기밀 유출이 가장 많은 것으로 나타났다. 유출 방법은 인력 매수를 통한 방식이 77건으로 가장 많았고, 무단 반출, 내부 공모, 공동 연구, 위장 합작이 순으로 인력 매수 부분이 가장 많았다. 분야별로는 전기 전자, 정보통신, 자동차, 바이오, 조선, 기타 등 전기 전자 부분이 가장 많았고 기술유출 동기는 개인영리, 금전 유혹, 처우불만, 인사·불만, 비리연루, 기타 순으로 나타났다[1].

이와 같이, 내부 인력에 의한 핵심기밀 유출이 대부분이고 각 기업은 인원 보안에 대한 대책 마련과 관리 설정이 필요함에 따라 각 기업에서는 핵심기술 유출 방지 시스템 구축에만 집중할 것이 아니라 인원 보안의 관리 설정에 중점을 두어 반복되는 핵심기밀 유출을 최소화 할 수 있도록 대책을 마련해야 한다.

본 논문은 기업의 핵심기밀 유출자 가능성 분류에 대해 제안한다. 본 제안 방법은 데이터 마이닝 방법 중 군집화 방법을 이용하여 중요 정보관리 부서 및 핵심 정보 관리 부서와 같은 주요 정보 부서에 자주 방문하거나 의심의 소지를 갖고 있는 사원들을 그룹핑 하여 관리한다. 이와 같이 그룹핑 함으로서 자주 접근하는 사원 및 자주 접근하지 않는 사원들을 분류하여 관리 할 수 있다.

본 논문이 제안하는 방법을 이용하면 기업에서의 핵심기밀 유출에 대한 피해를 최소화 할 수 있고 미연에 방지를 할 수 있을 뿐만 아니라 나아가 국가적인 피해 손실에 대해서도 최소화 할 수 있고 기업으로서는 효과적으로 내부 인원에 대해 관리할 수 있다.

본 제안 방법을 설명하기 위해서 본 논문의 구성은 다음과 같다. 제 2장에서는 핵심기밀 유출 방지 관련 연구에 대해서 설명하고, 제 3장에서는 본 논문에서 제안하는 군집화를 이용해 유출자 분류 방법을 설명 한다. 제 4장에서는 본 논문에 제안하는 방법의 안정성과 효율성을 알아보고, 제 5장에서 결론을 맺는다.

2. 관련 연구

2.1 유출방지체계 구축

이기혁 외 1명은 *내부정보 유출 징후 분석을 통한 유출방지체계 구축에 관한 연구*에서는 심각도와 위험도를 통해 정상 행위와 비정상 행위자를 분류했다. 이와 같은 분류에 앞서 정책 개발 방법론을 통해 PPP(policy, People, Process) 관점에서 내부정보를 취급하는 핵심인력 및 보안통제 위반자 등 집중 관리 대상자를 중심으로, 내부정보 취급 업무 시 야기 될 수 있는 내부 정보 유출 위험을 탐지, 모니터링, 예방하고 지속적으로 관리할 수 있는 정책을 제시하였다[2].

1단계 : 단위 Rule 수립
2단계 : 단위 Rule 분류
3단계 : Rule 영역별 분류
4단계 : 시나리오 분석
5단계 : Forensic 분석

위 5가지의 PPP를 기반으로 내부정보 유출분석 정책개발을 수립 하였다.

2.2 유출 예방 기업내 컴퓨팅 환경 최적화

송성근 외 3명이 제안한 *정보기술 유출 예방을 위한 기업內 컴퓨팅 환경 최적화 방안 연구*에서는

기업에서 협력업체의 아웃소싱하는 방식이 많아짐에 따라 기업 내부 기밀에 쉽게 접근할 수 있고 정보 소스를 유출 할 수 있는 사례가 빈번함에 따라 외부 개발환경 영역, 내부 개발 환경을 구분하였다.

이 연구에서는 외부 개발자에 대한 보안정책을 적용하여

- 외부 개발자 준수사항
- 외부 개발자 센터 개발자 업무처리
- 보안 관리자의 업무처리
- 방화벽 및 컨텐츠 필터링 툴 등록 절차

와 같이 보안성 향상을 위해 제반 보안 정책을 제안하였다[3].

2.3 사용자 인증 기술

산업기술 보호 연구에서 가장 오래도록 연구되어 온 것이 사용자 인증 기술이다. 사용자 인증 기술은 물리적 인증 기술과 논리적 인증 기술로 나눌 수 있으며, 사람의 바이오 정보를 이용한 바이오 인증 기술과, RFID와 같은 센서를 이용한 센서인증 기술이 대표적이다[4].

사용자 인증은 기업 내부 보안의 핵심적인 기술이며 기업 출입 통과시 제일 먼저 확인하는 것이 사용자 인증인 만큼 쉽게 접하면서도 가장 민감한 영역이다.

사용자 인증 기술은 다음과 같이 분류 할 수 있다.

가. 물리적 인증기술

- 지문인식을 이용한 사용자 인증
- 얼굴인식을 이용한 사용자 인증
- 센서정보를 이용한 사용자 인증

나. 논리적 인증기술

- 암호기술을 이용한 논리적 인증
- 바이오 정보를 이용한 논리적 인증

위와 같은 방법으로 물리적 인증 및 논리적 인증으로 사용자 인증 기술을 열거 할 수 있다.

3. 제안 방법

본 논문이 제안하는 방법은 데이터 마이닝 알고리즘을 이용하여 유출자 행위의 기본정보를 토대로 그들 간의 관계를 데이터 내에서 발견된 정보들을 근간으로 객체들을 그룹으로 만드는 것이다.

유출자 행위자를 구별하기 위해선 기업의 사원들에 속성이 필요하며 이 속성들은 입사 때부터 쌓여진 데이터를 근간으로 둔다. 개인의 속성을 정의하기 위해서 다음과 같이 정의 하였다.

3.1 개인정보 속성 정의

사원에 대한 개인정보 수집은 개인정보보호에 위배되지 않게 이름과 주민번호를 제외한 최대한의 속성으로 <표 1>과 같이 정의하였다. 데이터의 척도를 나타내기 위해선 수치나 기호가 필요하다. 따라서 데이터 타입이 String을 제외한 나머지 속성의 값은 그대로 사용하고 데이터 타입이 String인 속성은 별도의 수치로 나타내기로 한다.

〈표 1〉 개인정보 속성 정의

사원	데이터 타입	표기
성별	Integer	Sex
나이	Integer	Age
근속기간	Integer	Date
부서이름	String	Dept
접근횟수	Integer	Acc
비밀등급	String	Lev
이직횟수	Integer	Mov

3.2 제안 알고리즘

3.2.1 의사 결정 트리(Decision Tree)

의사결정 트리는 일련의 질문들을 통해 해당 레코드의 클래스 레이블에 대한 결론에 도달할 때까

지 후속 질문이 이어진다. 일련의 질문들과 가능한 대답들은 노드(node)와 방향 간선(directed edge)으로 이루어진 계층적인 구조인 의사결정 트리의 형태로 구성된다.

의사결정 알고리즘 중 헌트 알고리즘은 의사결정 트리 귀납 알고리즘의 기초가 되고 있다.

3.2.1 군집화(Clustering)

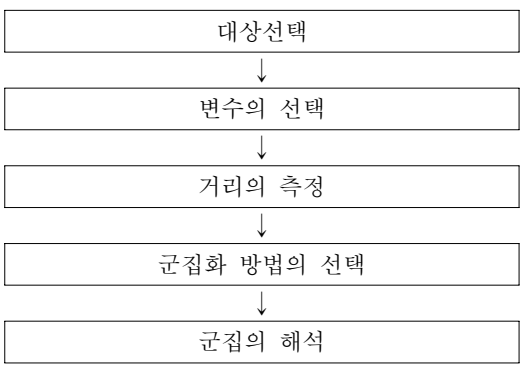
군집화 분석이란 객체와 그들 간의 관계를 기술하는 데이터 내에서 발견된 정보들을 근간으로 데이터 객체들을 그룹으로 만든 것이다. 같은 그룹에 속한 객체들의 유사성은 높이고, 다른 그룹의 객체들과의 차이점을 벌리는 것이 목적이며, 그룹 내 객체 간의 유사성과 그룹 간의 차이점을 높일수록 군집화는 점점 뚜렷해진다[5].

(1) 집괴법

각 객체를 하나의 군집으로 간주함을 시작으로 유사한 객체들을 묶어 군집으로 만들고 다시 유사한 군집들을 묶어 새로운 군집을 만들어 나가는 과정을 전체의 객체들이 하나의 군집이 되기까지 반복한 후 어떤 규칙에 의하여 최종적인 군집결과를 제공하는 방법

(2) 분리법

전체의 객체를 하나의 군집으로 간주함을 시작으로 유사성이 떨어지는 객체들을 분리시켜 다른 군집으로 만들어 나가는 과정을 각 객체가 하나의 군집이 될 때까지 반복 한 후 어떤 규칙에 의하여 최종적인 군집결과를 제공하는 것이다.



(그림 1) 군집분석의 수행절차

4. 성능 평가

본 논문이 제안하는 군집화 방법을 이용하여 기업에서의 기밀 유출자 행위 분석을 토대로 유출 가능성 있는 그룹을 도출 하는 것으로 성능 평가를 맺는다.

4.1 데이터 산출

개인정보 수집은 경기대학교 산업기술보호특화센터 직원 및 연구원의 동의를 얻어 3년간의 데이터를 토대로 수집 되었으며 실험 데이터로서 사용하였다.

4.2 실험 데이터

<표 2>와 같이 ID는 순차적 번호를 의미하며 특정한 사람을 지칭하지 않는다, 성별은 남자와 여자를 구분하고 부서는 총 네 가지의 부서로 나뉘어지며 숫자로 표기 하였다. 비밀등급은 비밀접근구

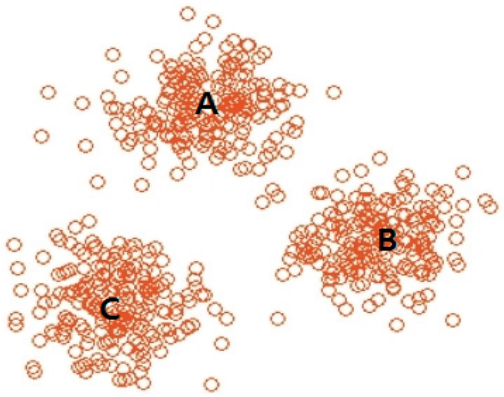
〈표 2〉 실험데이터 표

속성	데이터 척도(수치화)
ID(번호)	1, 2, 3, 4, …, n
Sex(성별)	남자 : 1, 여자 : 2
Age(나이)	1, 2, 3, 4, …, n
Date(근속기간)	1, 2, 3, 4, …, n
Dept(부서)	연구위원부 : 1 전략기획부 : 2 연구개발부 : 3 사업지원부 : 4
Acc(접근횟수)	1, 2, 3, 4, …, n
Lev(비밀등급)	1급 : 1 2급 : 2 3급 : 3
Mov(이직횟수)	1, 2, 3, 4, …, n
CLASS(위험군)	상 : 3 중 : 2 하 : 1

역에 따른 비밀 등급이 나뉘어지고 접근시간은 위험군은 군집화로 인한 위험도 분류에 의해 그룹이 나뉘어진다. 마지막으로 나이, 근속기간, 접근횟수, 이직 횟수는 무한한 수를 갖을 수 있다.

4.3 알고리즘 성능 분석

지도학습(Supervised Learning) 데이터 마이닝 기법 중 의사 결정 트리(Decision Tree)는 어떤 집합을 여러 개의 집합으로 쪼개어 각각을 어떤 특정한 성질을 가지는 클래스로 구분하는 분류화 방법의 한 부류이다. 또한 지도학습 데이터 마이닝은 공격 트래픽 학습 과정과 공격 탐지 과정이 확연히 구분되어 실시간으로 네트워크 상태를 반영하는 점진적 학습의 수행이 어렵고, 레이블링된 데이터가 필요하다는 제한이 있다. 따라서 지도학습의 이러한 문제점들을 해결하고자 비지도 학습(Unsupervised Learning) 데이터 마이닝 기법 중 군집화를 이용한 모델을 제안하였다[6].



(그림 2) 군집화에 따른 위험군 분류

트레이닝 데이터는 관제실 및 주요부서 접근 횟수에 따라 분포도 성향이 틀린 것으로 나타났고 그림과 같이 A, B, C와 같이 상위 위험군에 속하는 A 클래스에는 전략기획부 및 연구개발실 사원들이 대부분 속해 있었고 비밀접근구역이나, 일반적이지 않은 부서에 많이 접근하는 성향을 보였다.

〈표 3〉 유사도 측정 결과

군집반복 횟수	이전 유사도	최근 유사도
1	5.09E+06	3.59E+06
2	3.59E+06	3.16E+06
3	3.16E+06	2.61E+06
4	2.61E+06	1.39E+06
5	1.39E+06	1.29E+06
6	1.29E+06	1.21E+06
7	1.21E+06	1.21E+06

<표 3>은 최적의 군집을 얻기 위해 여러 반복을 통하여 군집간의 유사도를 축소하기 위해 나타난 유사도 측정 값이다. (그림 2) 결과를 토대로 데이터의 강한 응집도와 낮은 결합도를 도출할 수 있으며 시간의 흐름에 따라 작업 순서가 정렬되는 응집 관계가 형성되고 유사한 성격의 개체들이 그룹지어 모여지는 모습을 볼 수 있다. 또한 낮의 결합도로 인해 각 그룹간의 의존도가 낮아지기 때문에 비밀접근구역에 자주 접근하는 그룹과 그렇지 않은 그룹의 성향을 구분 지을 수 있으며 위험도 그룹 분류도 가능해 지는 것을 결과를 통해 알 수 있다.

발생함에 따라 인적보안에 대한 연구를 데이터 마이닝 기법을 통해 유출자를 분류 할 수 있는 방법을 제안하였다.

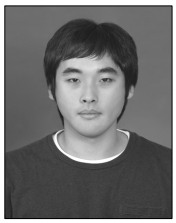
본 논문에서 어려웠던 점은 개인정보 수집에 있어서 대규모 기업의 과거 기밀 정보 유출 실사례를 기반으로 한 트레이닝 데이터의 수집이 어려웠고 다소 신뢰성이 떨어지는 문제점이 있다.

향후 연구과제로는 다양한 군집화 분석을 통하여 다각적인 시각과 실제 트레이닝 데이터의 비교 연구를 통해 좀 더 신뢰성 있는 연구 결과를 도출하는 것이 필요하다.

참 고 문 헌

[1] 국정원 산업기밀보호센터, 2010.
[2] 이기혁, 이철규, “내부정보 유출 징후 분석을 통

한 유출방지체계 구축에 관한 연구", 정보보학회지, 제19권, 제3호, 2009.
[3] 송선근, 박지숙, 우재현, 임종인, "정보기술 유출 예방을 위한 기업內 컴퓨팅 환경 최적화 방안 연구", 정보보호학회지, 제18권, 제6호, 2008.
[4] 이대성, 김재성, 김귀남, "정보 유출 방지 연구 기술 동향", 정보보호학회지, 제20호, 제1권, 2010.
[5] http://en.wikipedia.org/wiki/Cluster_analysis.
[6] Leonid Portnoy, "Intrusion detection with unlabeled data using clustering", Undergraduate Thesis, Columbia University, 2000.



허 승 표

2009년 남서울대학교 컴퓨터학과(공학사)
2010년 현재 경기대학교 산업보안학과 산업과정

이 대 성

1999년 인하대학교 전자계산공학과(학사)
2001년 인하대학교 전자계산공학과(석사)
2005년 아안픽춰스 VR 연구소선임 연구원
2008년 인하대학교 정보공학과(박사)
2010년 현재 경기대학교 산업기술보호특화센터 연구교수

김 귀 남

미국 캔자스대학(학사)
미국 콜로라도주립대학(석사)
미국 콜로라도주립대학(박사)
현재 경기대학교 산업보안학과 교수