Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

# Data-mining Based Anomaly Detection in Document Management System

**Hyun-Song Jang***

*Department of Business Administration, Seoul School of Integrated Sciences & Technologies*

**ABSTRACT**

The problem of security management in document management system is that it is even harder to detect an authenticator who uses the system anomalously, although it is as much as possible to prevent and detect the movement of exposing information or destroying the system as a malicious purpose by means of passing through physical and logical protective devices of the system. To solve this problem, an usage status of a contents has been collected within a certain period of time, which is actually used at a specific site and the behavior patterns of a user based on data mining technique which has been analyzed. The main variable deciding a user's pattern has been analyzed and the user has been clustering according to the result. A single-host based on anomaly detection model was designed by techniques of K-Means and Self-Organizing Maps being used for clustering. This model detects a case where a specific user deviates the existing result of clustering in comparison of the result of experiment as anomaly. The content usage pattern of a user was utilized for the designed detection model while the existing detection model of anomaly was established through analysis of usage pattern of command or data packet. In this study, it is to discuss data mining-based anomaly detection model in which couldn't be solved out by the existing method, detection of a user's intentional exposure of information, and the result of the experiment.



**KEYWORDS:** Data mining, Anormaly detection, Security management, Document management

**ARTICLE INFO:** Received 27 July 2015, Revised 14 August 2015, Accepted 14 August 2015.

*Corresponding author is with the Department of Business Administration, Seoul School of Integrated Sciences & Technologies, 46 Ewhayeodae 2-gil, Seodaemun-gu, Seoul 120-808, KOREA.

*E-mail address*: hyunsong.jang@stud.assist.ac.kr

## 1. 서 론

국정원 산업기밀보호센터 통계자료에 따르면 2003년부터 2014년까지 국내 발생 산업스파이 사건 현황 중 해외 기술유출 적발 건은 총 438건으로 국가 기술경쟁력이 높아질수록 산업스파이도 증가하는 것으로 나타났다. 기술유출로 약 50조원의 연평균 피해금액이 발생하는 것으로 추정되며, 대부분이 전직, 현직 직원 등 내부유출이 79.9%를 차지하고 있으며, 기술의 가치를 잘 알고 기술 접근 권한이 있는 내부 사용자의 불법적인 유출이 대다수인 것으로 파악되었다<그림1>.

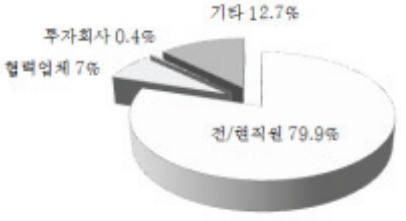


그림 1. 해킹 유형 및 발생 건수
Figure 1. Hacking type and cases

이처럼 외부에서의 침입보다는 내부의 인증된 사용자의 비정상적인 사용이 보다 큰 문제임에도 불구하고, 기존의 침입 탐지 시스템이 외부의 침입이나 네트워크상의 데이터 분석을 통한 정보의 유출에 많이 집중되는 이유는 인증된 사용자의 비정상적인 행위를 탐지하기 어렵기 때문이다[1].

기존의 비정상 행위 탐지 시스템은 시스템 명령어 사용에 대한 패턴이나 시스템 자원의 사용률을 통계적인 기법을 사용하여 탐지하였으나, 이러한 탐지 방법으로는 콘텐츠 관리 시스템 같은 상용 프로그램 기반의 대규모 시스템에서는 적용에 어려움이 있고, 탐지 결과가 정확하지 않았다.

문제를 해결하고자 내부의 인증된 사용자의 비정상 행위를 탐지하기 위한 방법으로 인증된 사용자가 정보 획득을 위해 조회한 콘텐츠, 그리고 해당 콘텐츠와 관련된 제품 종류나 특정 기술을 파악하여 사용자 별로 특정 유형의 패턴이 있다고 가정하고 이를 데이터마이닝 기법을 사용하여 클러스터링 하였다[2]. 이렇게 사용자의 콘텐츠 사용 패턴은 인증된 사용자가 비정상적인 행위를 시도할 경우 이를 탐지할 수 있는 기준이 되었다.

본 논문에서는 반도체 업계에서 선두의 기술력을 보유하고 있는 기업의 문서관리 시스템을 대상으로 일정 기간 콘텐츠 중심의 사용 현황을 분석하여 사용 패턴을 결정지을 수 있는 요인들을 설정하였다[3]. 설정된 인자를 사용하여 조직 내부의 정상적인 사용자의 비정상적인 사용 행위를 탐지할 수 있는 모델을 정립하고 정립된 모델을 사용한 실험 및 결과 제시하였으며, 본 모델을 기초로 내부 사용자의 악의적인 의도의 정보 유출을 탐지할 수 있는 시스템을 제안하고자 한다.

## 2. 관련 연구

J.P Anderson은 1980년 침입 탐지의 개념을 도입하면서 허가되지 않은 사용자가 정보를 조회하거나 조작하고, 시스템을 신뢰성이 없거나 사용할 수 없도록 하는 것을 침입이라고 정의하고, 침입자는 외부로부터의 침입과 내부의 침입으로 구분하였다[4].

S.E Smaha 는 침입 유형과 탐지할 수 있는 방법을 6가지로 분류하였으나[5], 일반적으로 침입 탐지는 비정상 행위 탐지(Anomaly Detection)와 오용 탐지(Misuse Detection)의 두 가지로 분류한다.

비정상 행위 탐지는 모든 침입 행위에는 비정상적인 행위가 필요하다는 가정에서 출발한다. 정상적인 시스템 사용에 대한 프로파일 상태를 유지하면 비정상적인 행위에 대한 탐지가 가능하다<그림 2>.

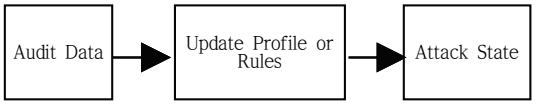


그림 2. 일반적인 비정상 행위 탐지 시스템
Figure 2. General type of anomaly detection system

오용 탐지는 알려진 취약성에 대한 공격 유형을 사전에 알고, 공격이 시도될 때 이를 탐지하는 방식이다. 비정상 행위 탐지가 침입으로 여겨지는 행위를 탐지한다면 오용 탐지는 바이러스 백신 프로그램처럼 명백한 침입을 탐지한다[6].

## 3. 실험시스템 개요 및 전처리

실험에 사용된 시스템은 반도체 업계에서 세계적인 경쟁력을 가지고 있는 회사의 문서 관리 시스템으로 약 300만 건의 제품 개발 및 생산과 관련된 주요 문서를 전 세계 인증된 사용자에게 연중 무정지로 제공하는 시스템이다. 약 40,000 명의 인증된 사용자가 등록되어 있으며, 월 평균 약 30,000건의 콘텐츠가 12,000 여명의 사용자에 의해 조회되고 7,000 여건의 문서가 신규로 생성된다.

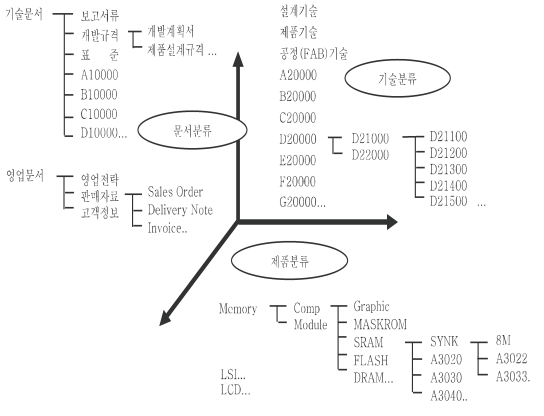


그림 3. 콘텐츠 구성 현황
Figure 3. Contents type in dataset

각 콘텐츠는 문서 종류, 제품 종류, 기술 분류에 의해 입체적인 분류 체계를 가지고 있으며, 3단계의 보안 레벨이 작성자에 의해 명시된다<그림 3>.

사용자는 이들 분류 및 보안 레벨을 선택적으로 사용하여 원하는 콘텐츠를 조회한다. 또한 사용자는 부서 및 경력에 따른 300 여개의 단위 부서그룹과 16 여개의 직무그룹으로 그룹화 되어 있다.

실험 시스템 현황을 토대로 특정 사용자는 특정 콘텐츠 분류 체계 내의 콘텐츠를 일정 기간 내에 동일한 패턴을 지닌 사용한다는 가정을 생각할 수 있다. 사용자는 이미 부서 및 직무에 따라 그룹화 되어 있으나, 동일한 부서나 직무의 사용자일지라도 시스템 내의 콘텐츠를 사용하는 패턴은 다를 수 있으므로 이를 파악하여 개인별 시스템 사용 패턴으로 설정하고 시스템 사용 시 설정된 패턴과의 유사성 여부를 판단하여 비정상 행위 여부를 판단할 수 있다[7].

실험에 사용된 관측 데이터 구조는 <표 1>과 같다.

표 1. 원시 데이터 테이블
Table 1. Data table layout

| 필드명 | 데이터 타입 | 설명 |
|---|---|---|
| Emp_ID | Character | 사용자 ID |
| Content_type_1 | Decimal | 기술보고서 |
| Content_type_2 | Decimal | 개발규격 |
| Content_type_3 | Decimal | 기술논문 |
| Content_type_4 | Decimal | 제안 |
| Content_type_5 | Decimal | 특허 |
| Content_type_6 | Decimal | 컨텐츠 종류6 |
| Content_type_7 | Decimal | 컨텐츠 종류7 |
| Content_type_8 | Decimal | 컨텐츠 종류8 |
| Product_code_1 | Decimal | Memory관련 컨텐츠 |
| Product_code_2 | Decimal | A Project 관련 컨텐츠 |
| Product_code_3 | Decimal | LSI 관련 컨텐츠 |
| Technical_code_1 | Decimal | Design관련 컨텐츠 |
| Technical_code_2 | Decimal | Process 기술 컨텐츠 |
| Technical_code_3 | Decimal | 조립 기술 컨텐츠 |
| Technical_code_4 | Decimal | 테스트 기술 컨텐츠 |
| Technical_code_5 | Decimal | 공통 기술 컨텐츠 |

데이터 마이닝[8]을 위한 데이터 전처리는 SAS Program을 사용하여 가공하였으며, SAS Enterprise Miner의 클러스터링 모듈의 k-means 와 SOM(Self Organizing Map) 클러스터링을 사용하였다[9].

각 사용자가 조회한 콘텐츠의 종류, 관련 제품코드, 기술코드로 레코드가 구성되어 있으며, 각 값은 조회한 횟수 이다.

이 외에도 콘텐츠의 보안 등급이나 IP Address 등 사용자 패턴을 결정지을 수 있는 요소가 추가적으로 있으나, 클러스터링을 콘텐츠 사용 패턴에 집중하기 위하여 실험에서는 배제하였다[10].

군집 분석을 수행하기 위해서는 원시 데이터 세트 결측치가 포함될 경우 그 자체로 군집이 되거나 다른 군집을 왜곡시키므로 결측치나 이상치에 대해 전처리를 해 주어야 한다. 수집된 관측 데이터 77만여 건 중, 결측치 데이터를 포함하는 5만4천여 건의 데이터는 반영하지 아니하였다.

사용자 정보가 일반 사용자가 아닌 시스템 사용자일 경우 사용자 정보가 달라 결측치로 가정하여 배제하였고, 또는 동일 사용자에 대하여 월 평균 4회 이하의 조회 건수에 대한 데이터는 시스템을 자주 사용하지 않는 사용자로 파악되었고, 이는 전체 클러스터링을 왜곡할 수 있어 이상치로 처리하였다.

또한 업무상 조사나 감사 등의 목적으로 일시적으로 사용된 데이터의 경우, 동일 사용자 기존 패턴과 다르게 월 평균 60회 이상 발생하는 등 특수 목적에 의해 시스템을 전용으로 사용하는 경우도 있었으며 이 역시 전체 클러스터링을 왜곡할 수 있는 가능성이 있어 이상치로 가정하고 배제하였다.

원시 데이터로부터 효율적인 클러스터링을 위해 전처리가 완료된 실험 데이터 세트를 SAS Enterprise Miner에서 조회한 화면은 <그림 4>와 같다.

| Name | Model R | Measure | Type | Format | Inform | Variable Label |
|---|---|---|---|---|---|---|
| EMP_ID | Id | nominal | char | $255. | $255. | emp_id |
| CONTENT_TYPE_1 | input | binary | num | BEST12. | 12. | content_type_1 |
| CONTENT_TYPE_2 | input | binary | num | BEST12. | 12. | content_type_2 |
| CONTENT_TYPE_3 | input | binary | num | BEST12. | 12. | content_type_3 |
| CONTENT_TYPE_4 | input | binary | num | BEST12. | 12. | content_type_4 |
| CONTENT_TYPE_5 | input | binary | num | BEST12. | 12. | content_type_5 |
| CONTENT_TYPE_6 | input | binary | num | BEST12. | 12. | content_type_6 |
| CONTENT_TYPE_7 | input | binary | num | BEST12. | 12. | content_type_7 |
| CONTENT_TYPE_8 | input | binary | num | BEST12. | 12. | content_type_8 |
| PRODUCT_CODE_1 | input | binary | num | BEST12. | 12. | product_code_1 |
| PRODUCT_CODE_2 | input | binary | num | BEST12. | 12. | product_code_2 |
| PRODUCT_CODE_3 | input | binary | num | BEST12. | 12. | product_code_3 |
| TECHNICAL_CODE_1 | input | binary | num | BEST12. | 12. | technical_code_1 |
| TECHNICAL_CODE_2 | input | binary | num | BEST12. | 12. | technical_code_2 |
| TECHNICAL_CODE_3 | input | binary | num | BEST12. | 12. | technical_code_3 |
| TECHNICAL_CODE_4 | input | binary | num | BEST12. | 12. | technical_code_4 |
| TECHNICAL_CODE_5 | input | binary | num | BEST12. | 12. | technical_code_5 |

그림 4. 컨텐츠 구성 현황
Figure 4. Contents type in dataset

관측치 사이의 거리를 측정하는 변수로는 이진형(Binary), 범주형(Catergorical), 순서형(Ordinal), 구간형(Interval) 이 있는데, 분산이 큰 변수들은 분산이 작은 변수들보다는 클러스터링을 형성하는데 더 많은 영향을 주므로 사전에 중요도가 정의된 경우를 제외하고는 모든 변수가 동일한 중요도를 갖도록 표준화하는 작업이 필요하다. 본 실험에서는 컨텐츠의 사용 패턴을 구분하기 위한 컨텐츠 종류, 제품, 기술의 세가지 변수에 대한 조회수인 구간형 데이터를 가지고 실험하며 각 변수에 대한 중요도는 동일하다고 가정한다. 위 실험 데이터를 가지고 K-means와 SOM을 사용하여 클러스터링한다.

## 4. 실험 및 결과

데이터 전처리 후 K-means와 SOM 각각의 방법으로 클러스터링한 후 생성된 의사결정나무를 사용하여 정상/비정상 데이터를 사용하여 다름과 같은 과정으로 실험되었다<그림 5>.

비교실험에 사용된 SAS Enterprise Miner 4.0 은 K-means에서 클러스터의 개수를 자동으로 생성하거나 수동으로 정의하는 것이 모두 가능하다.

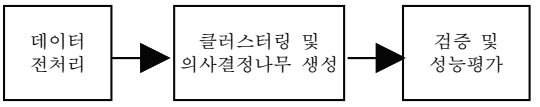


그림 5. 실험과정
Figure 5. Experimental methods & procedure

자동으로 클러스터의 개수를 생성하는 것은 프로그램 내의 Cubic Clustering Criterion 그래프를 사용하여 결정되는데, 본 실험에서는 K-means 방법과 SOM 방법에 대한 결과를 비교하기 위하여 자동으로 생성된 K-means클러스터의 개수를 SOM에도 적용하였다. 본 실험의 경우, 자동으로 K-means에서 자동으로 클러스터링 된 개수는 4개이며 SOM의 경우 그 결과와 비교하기 위해 같은 4개로 클러스터로 정의하였다.

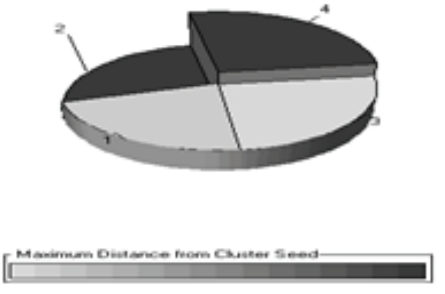


그림 6. K-means 클러스터링 결과
Figure 6. Result of K-means clustering

〈그림6〉 는 K-means 클러스터링 결과이며 Pie Chart는 조각의 폭, 높이, 색으로 표현되는데, 이는 각 클러스터의 세가지 통계값을 보여준다. 각 클러스터의 높이는 해당 클러스터에 포함된 관측치의 수(Frequency)를 의미하며, 각 클러스터의 폭은 그 군집의 표준편차를, 색상은 군집에 속한 관측치와 중심 사이 거리의 최대값인 클러스터 반경(Radius)을 의미한다.

SOM에 의한 클러스터링 결과이며 수동으로 Map의 행, 열 수를 지정해줘야 하는 점이 K-means 클러스터링과 다르다. 설정된 맵의 크기가 너무 작으면 자료에 존재하는 비선형성의 관계를 표현하기 어렵고, 반대로 너무 크면 분석에 많은 시간이 소요되고 관찰치가 하나도 포함되지 않은 군집이 발생하여 분석 결과에 대해 그릇된 해석을 할 수 있다〈그림 7〉.

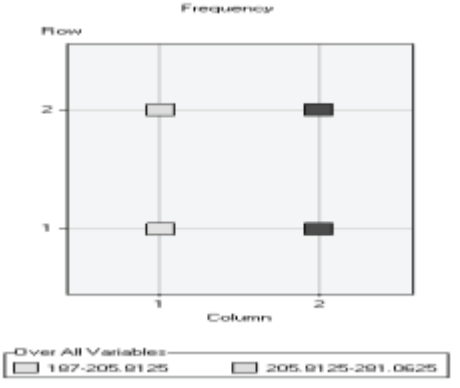


그림 7. SOM 클러스터링 결과
Figure 7. Result of SOM clustering

실험 결과, K-means에 의한 클러스터는 4개로 지정되었으며 각 클러스터에 포함된 데이터는 다른 클러스터와 중복된 데이터가 없었고, SOM의 경우, K-means 와 비교하기 위하여 행, 열을 2, 2로 지정하여 4개의 클러스터를 구성하였으나, 데이터 범위가 중복되는 클러스터가 존재하였다.

클러스터간 거리 그래프는 각 클러스터에 대한 위치와 관계를 그래프로 보여주는 것으로 별표는 클러스터의 중심을 의미하고 원의 크기는 군집의 범위를 의미한다. K-means의 경우, 원이 다른 클러스터의 범위와 중복된 것은 클러스터에 속한 데이터가 중복되었다는 의미는 아니며, 데이터의 퍼져있는 범위를 의미하는 것이다〈그림 8〉.

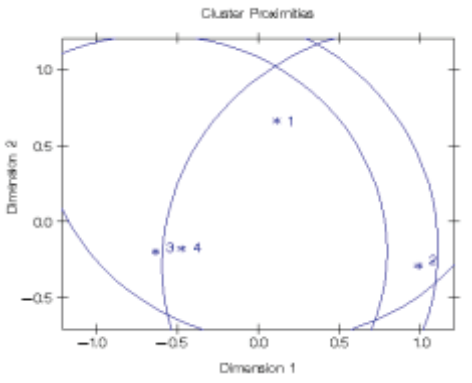


그림 8. K-means 클러스터간 거리그래프
Figure 8. Distance chart of K-means clustering

SOM 클러스터링의 경우, 클러스터의 범위를 나타내는 원이 없는데, 이것은 SOM으로 처리했을 경우, 데이터의 범위가 그래프에 표시하기 어려울 정도의 크기이기 때문이다. 이런 그래프일지라도 각 클러스터는 중복된 데이터를 가지는 것은 아니며 단지 데이터를 클러스터링 할 때 클러스터링 방법에 따른 백분율에 의해 클러스터를 지정해야 한다 <그림 9>.

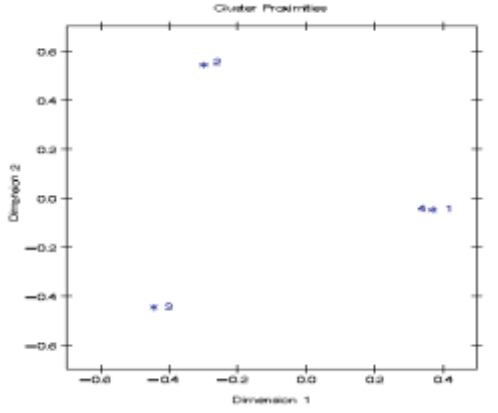


그림 9. SOM 클러스터간 거리그래프
Figure 9. Distance chart of SOM clustering

본 실험에서 K-means 는 중복되는 범위가 없이 클러스터링 되었으나, SOM 의 경우에는 각 클러스터 간 중복되는 데이터의 범위를 가지고 있다. <그림 10>은 K-means에 의한 의사결정나무(Cluster Decision Tree)의 일부이며 모든 관측 데이터가 이 의사 결정 나무에 의해 클러스터링 되어있음을 확인할 수 있다.

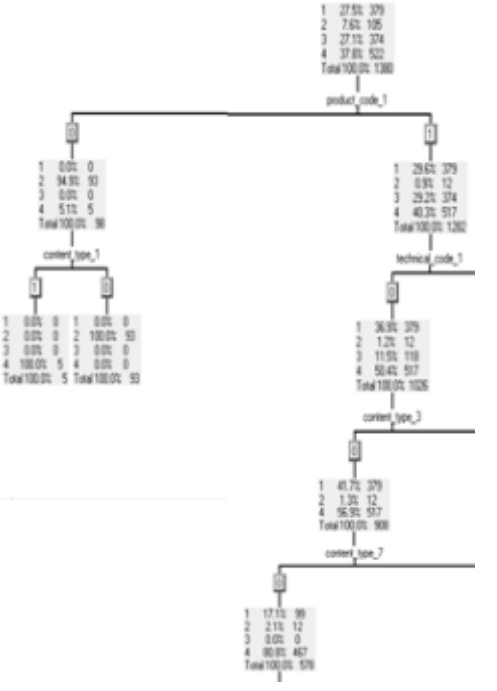


그림 10. K-means 의사결정나무
Figure 10. Decision tree(K-means)

생성된 의사 결정 나무는 IF-THEN 논리식으로 표현할 수 있으며 테스트 관측치를 본 논리식으로 수행하여 동일한 클러스터가 되었을 경우, 실험 데이터에 대한 클러스터가 사용자 행위에 대한 패턴을 잘 표현하였다고 판단한다. 결과를 만족하였을 때 이를 비정상 행위를 판단하는 기준으로 삼을 수 있다<그림 11>.

```
IF content_type_1 EQUALS 1
AND product_code_1 EQUALS 0
THEN
  NODE   :    4
  N      :    5
  1      :  0.0%
  2      :  0.0%
  3      :  0.0%
  4      : 100.0%

IF content_type_1 EQUALS 0
AND product_code_1 EQUALS 0
THEN
  NODE   :    5
  N      :   93
  1      :  0.0%
  2      : 100.0%
  3      :  0.0%
  4      :  0.0%

IF technical_code_1 EQUALS 1
AND product_code_1 EQUALS 1
THEN
  NODE   :    7
  N      :  256
  1      :  0.0%
  2      :  0.0%
  3      : 100.0%
  4      :  0.0%

...
```

그림 11. K-means 의사결정나무 IF-THEN 논리식
Figure 11. IF-THEN logical expression(K-means)

생성된 의사 결정 나무와 별도의 검증용 데이터를 사용하여 실험용 데이터의 클러스터와 검증용 데이터가 얼마나 일치하는지 실험하였다. 검증용 데이터는 실험용 데이터 세트에 포함되어있는 사용자 ID 중 실험용 데이터 전후 3개월 내 기간의 데이터에서 무작위로 100 건을 추출하여 사용하였다.

정상행위는 추출된 100 건의 데이터가 생성된 클러스터와 일치 여부를 검사하기 위해 의사결정나무 논리식을 적용하여 계산하고, 비정상 행위는 추출된 데이터를 임의로 변경하여 기존 사용자 패턴과 상이한 패턴을 가지도록 가공하였다. 예를 들면 평소에는 DRAM제품의 Test기술과 관련된 기술보고서를 주로 보던 사용자의 데이터를 LSI제품의 설계 관련 도면을 조회하는 내용으로 변경하였다.

실험 결과, 정상행위에 대해서는 K-means 98%, SOM 95%로, 비정상 행위는 K-means 와 SOM 모두 100% 탐지할 수 있었다<표 2>.

표 2. 클러스터링 성능평가
Table 2. Clustering result validation

| 구분 | | 실험데이타 | 일치/불일치 | 탐지율 |
|---|---|---|---|---|
| 정상 행위 | K-means | 100건 | 98/2 | 98% |
| | SOM | 100 건 | 95/5 | 95% |
| 비정상 행위 | K-means | 100건 | 100/0 | 100% |
| | SOM | 100 건 | 100/0 | 100% |

## 5. 결 론

실험 결과를 통하여, 인증된 사용자의 비정상적인 행위는 클러스터링에 의해 탐지될 수 있음을 확인할 수 있었으며, 정상적인 행위에 대한 판단은 약간의 오차 범위를 보여 정상적인 행위도 비정상적인 행위로 판단될 수 있음을 확인하였다.

비정상데이터를 실제 발생한 데이터로 실험하는 것이 가장 이상적이겠으나, 해당기간 동안 내부 사용자에 의한 정보유출이 없어 데이터 확보가 불가능했기에 내부 사용자가 기존의 시스템 사용 패턴을 벗어나서 사용한다는 상황을 전제로 가상데이터를 만들어서 실험하였으며, 이에 연구의 한계가 있다고 하겠으나 시나리오 기반 가상 데이터를 적용한 결과를 볼 때, 실제 상황에서도 무리없이 검출할 수 있을 것으로 예상한다.

본 논문에서는 기업의 중요한 지적 자산을 보호하기 위한 방법으로 내부 인증된 사용자의 비정상 행위를 탐지할 수 있는 방법으로 데이터마이닝 기반의 비정상 행위 탐지 모델을 제안하였다.

콘텐츠 사용에 따른 사용자 패턴을 인식하기 위하여 콘텐츠의 종류와 해당 콘텐츠와 관련된 제품 및 기술의 입체적인 분류 체계로 구성하고, 이를 중심으로 사용자 별 콘텐츠 사용 이력을 데이터화하여 전처리하였으며, K-mean와 SOM클러스터링 방법을 적용하여 실험한 결과, 비정상 행위의 탐지

가 가능하다는 결론을 얻었다.

본 논문의 연구내용을 기반으로 인증된 사용자에 대한 비정상행위를 실시간으로 탐지하는 시스템 구축에 관한 연구를 향후 연구 과제로 제시한다. 향후 제시된 연구 과제를 위해 보다 효과적인 사용자 패턴을 추출하기 위한 추가적인 연구와 많은 시스템 자원을 필요로 하는 클러스터링의 효율적인 운영 방법, 그리고 탐지 결과에 따른 자동화된 업무 프로세스 정의와 구현 방법을 연구할 예정이다.

## References


[1] S. Noel, D. Wijesekera, and C. Youman, *Modern intrusion detection, data mining, and degrees of attack guilt, in applications of data mining*, In Computer Security Kluwer Academic Publisher, Boston, 2002.

[2] Klaus Julisch. *Data mining for intrusion detection: A critical review*, In Applications of Data Mining in Computer Security. Kluwer Academic Publisher, Boston, 2002.

[3] J.F. Roddick, and B. G. Lees, *Paradigms for spatial and spatio-temporal data mining, geographic data mining and knowledge discovery,* Taylor and Francis, London, 2001.

[4] J.P Anderson, *Computer security threat monitoring and surveillance*, Technical report, James P Anderson. Washington, 1980.

[5] Steven E Smaha, *Haystack: An intrusion detection system,* In Fourth Aerospace Computer Security Applications Conference, pp. 37-44, 1988.

[6] Kumar, Sandeep, *Classification and detection of computer intrusions*. Diss. Purdue University, 1995.

[7] R. Schalfoff, *Pattern recognition-statistical, structural and neural approach*, John Wiley & Sons, 1992.

[8] R. Agrawal, T. Imielinski, and A. Swami, *Database mining: A performance perspective*, IEEE Transactions on Knowledge and Data Engineering, Special issue on Learning and Discovery in Knowledge-Based Databases, 9(6), pp. 914-925, 1993.

[9] J. Han, and M. Kamber. *Data mining: concepts and techniques*, Morgan Kaufmann, 2000.

[10] J.F. Roddick, and M. Spiliopoulou, *A bibliography of temporal, spatial and spatio-temporal data mining research*", SIGKDD Explorations 1, pp. 34-38, 1999.


## 데이터마이닝 기반 문서관리시스템 사용자 비정상행위 탐지에 관한 연구


**장현성**

*서울과학종합대학교 대학원 경영학과*



**요 약**

외부로부터의 물리적, 논리적 보안 장치를 통하여 시스템을 파괴하거나 정보를 유출하려는 행동은 일정 수준 예방과 탐지가 가능하나, 인증된 정상적인 사용자의 의도적 정보 유출 행위를 탐지하는 것은 어렵다.

문제 해결을 위하여 문서관리시스템의 사용 현황을 일정 기간 수집하여 사용자의 시스템 사용 패턴을 데이터 마이닝 기법을 사용하여 사용자의 시스템 사용 패턴을 결정하는 주요 변수를 분석하고 그에 따라 사용자를 클러스터링하였다. 클러스터링은 K-means 와 SOM 기법을 사용하였으며, 실험을 통하여 클러스터링 결과와 비교하여 특정 사용자가 기존 클러스터링 결과를 벗어날 경우, 이를 비정상적인 행위로 탐지하는

단일 호스트 기반의 비정상 행위 탐지 모델을 설계하였다. 본 연구를 통하여 기존의 방식으로는 해결할 수 없었던 정상적인 사용자의 의도적인 정보 유출을 탐지할 수 있는 데이터 마이닝 기반의 비정상 행위 탐지 모델과 실험 결과를 논하고자 한다.

**Hyun-Song Jang** received the B.S. from the Sungkyunkwan University, and M.S. in computer science from the Hanyang University. He is currently pursuing the Ph.D. in business administration at the Seoul School of Integrated Sciences & Technologies. He has been designing and establishing various information system for semiconductor, silicon wafer and LCD/OLED business.

*E-mail address*: hyunsong.jang@stud.assist.ac.kr