DOI: 10.3745/KIPSTA.2010.17A.4.181

# 네트워크 사용자에 대한 임베디드형 행동패턴 분석시스템

정 세 영[†] · 이 병 권[††]


## 요 약

본 연구에서는 데이터마이닝 기법을 이용하여 네트워크 사용자의 행동 패턴을 분석하는 시스템을 제안하고자 한다. 시스템 구현을 위해 네트워크 탭 장비를 사내 네트워크에 설치하여 패킷을 복제한다. 복제된 패킷은 고속의 메인 메모리DB를 통하여 데이터베이스에 저장된다. 저장된 데이터는 데이터 마이닝 기법을 이용하여 행동패턴을 분석하고 네트워크 관리자에게 실시간 보고한다. 또한 이기종간의 데이터를 공유하기 위해 표준의 XML 문서 교환 방식을 적용한다. 행동패턴 분석 시스템은 추후 저가 구현 및 쉬운 설치를 고려하여 임베디드형 셋톱박스 기반을 제안한다.


**키워드 : 행동패턴, 네트워크 탭, 데이터마이닝, 임베디드시스템, 표준-XML**

# An Action Pattern Analysis System of the Embedded Type about Network Users

Seyoung Jeong[†] · Byungkwon Lee[††]


## ABSTRACT

In this study, we suggest the system to analyze network users' action patterns by using Data-Mining Technique. We installed Network Tap to implement the analysis system of network action and copied the network packet. The copied packet is stored at the database through MainMemoryDB(MMDB) of the high-speed. The stored data analyze the users' action patterns by using Data-Mining Technique and then report the results to the network manager on real-time. Also, we applied the standard XML document exchange method to share the data between different systems. We propose this action pattern analysis system operated embedded type of SetToBox to install easily and support low price.


**Keywords : Action Patterns, Network Tap, Data-Mining, Embedded System, Standard XML**

## 1. 서 론

인터넷의 발달은 개인은 물론 기업의 업무에도 많은 영향을 주고 있다. 사내에서 근무하는 직원은 기본적으로 인터넷을 통해 업무에 도움이 되는 정보를 검색하여 활용하기도 하고, 홈쇼핑 및 각종 블로그를 통해 개인 활동을 하기도 한다. 이런 개인이 속한 대부분의 단체에는 업무의 종류로 부서를 구분한다. 부서 각각은 그 업무의 형태에 따라 일정한 패턴을 가지고 활동하게 된다. 물론 그 활동의 패턴은 인터넷 사용에도 반영 된다. 영업 부서에서 개발 부서가 사용하는 패턴의 움직임이 생긴다면 분명 이상 행위로 분류할 수 있다. 최악의 경우 사내의 기밀이 유출될 가능성도 예측할 수 있을 것이다. 일반적으로 운영부의 직원이 컴퓨터 프로그램을 작성에 필요한 정보를 담고 있는 사이트를 자주 방문한다거나, 개발부서에서 운영에 관한 사이트를 자주 방문하는 일은 특별한 경우가 아니면 거의 없을 것이다. 이러한 현상이 자주 발생된다면 회사 업무에 문제가 있거나 허가 받지 못한 사람이 타인의 컴퓨터를 사용하고 있는 경우이다. 이러한 움직임은 보안 장비나 바이러스 백신으로 잡아낼 수 없는 부분이다.

그리하여 본 연구에서는 네트워크 사용자의 웹 접근 빈도 및 시간을 데이터 마이닝기법과 메인메모리DB를 이용하여 일정한 행동패턴을 유추하려 한다. 마이닝된 데이터는 네트워크 사용자의 행동 패턴을 나타내며 이 패턴을 구현하기위한 시스템을 제시하고자 한다.

본 시스템의 구성은 망으로 들어오거나 나가는 모든 패킷을 네트워크-탭(Network-Tap) 장비로 복제(clone)하고, 복제된 패킷은 고속 메인메모리DB(MMDB)를 통하여 실시간

† 정 회 원 : 휴먼앤퓨처정보통신 선임연구원
†† 정 회 원 : 휴먼앤퓨처정보통신(주) 연구소장(교신저자)
논문접수 : 2010년 3월 9일
수 정 일 : 1차 2010년 5월 7일
심사완료 : 2010년 5월 7일

으로 저장된다. 수집된 데이터는 데이터마이닝 기법을 통하여 분석하고 마이닝 데이터를 생성한다. 패킷 데이터 수집 및 행동패턴분석 장비로 임베디드형 셋톱박스를 제안한다. 마이닝된 결과는 셋톱박스 어플리케이션에서 실시간으로 확인 가능하며, 표준 XML 포맷을 통하여 이기종인 웹서버에서도 확인 가능하도록 구현한다.

2장에서는 본 시스템에서 이용된 기술에 관한 관련연구를 기술하고, 3장에서는 시스템구성 및 구현에 대하여 설명한다. 그리고 4장에서 결론에 대하여 기술한다.

## 2. 관련연구

### 2.1 데이터마이닝 (Data Mining)

지식발견(KDD)은 데이터로부터 유용한 정보를 발견하는 프로세스 전 과정이고, 데이터마이닝은 지식발견 프로세스 중에서 데이터로부터 정보를 추출하기 위해 기법을 적용하는 특정단계라 정의한다[1]. 지금까지 알려진 데이터마이닝 기법들은 그 종류가 상당히 다양할 뿐만 아니라 지금도 새로운 기법들이 대학과 연구소를 통해 계속 소개되고 있다. 이 중에서도 일부만이 이론적인 검증을 거쳐 상품적 가치를 인정받고 있다. 그러나 데이터마이닝을 처음 접하는 초보자뿐만 아니라 실제 사례에 적용해본 경험이 있는 사람들에게도 자신들의 상황에 적합한 데이터마이닝 기법을 선택하는 작업은 그리 쉽지 않다. 이것은 데이터마이닝 작업 유형에 관계없이 가장 탁월한 성능을 제공하는 특정기법이 존재하는 것도 아니고 유사 기법이라고 하더라도 분석 대상이 되는 데이터의 특성이나 도출하고자 하는 정보의 성격에 따라 상이한 결과를 낳을 수 있기 때문이다[2].

본 연구에서는 수집된 데이터와 네트워크 사용자에 대한 패턴의 인과관계를 고려하여 통계적인 기법과 연관규칙 및 순차패턴 기법을 통하여 데이터마이닝 시스템을 구현하고자 한다.

#### 2.1.1 통계적 기법

통계적 기법은 대량의 데이터로부터 새롭고 의미 있는 정보를 추출하는 기술이고, 데이터로부터 정보를 뽑아내는 기능을 제공하기 때문에 넓은 의미에서 데이터마이닝 기법이라 해석할 수 있으며, 일부 데이터마이닝 관련 서적들도 이러한 해석에 따르고 있다. 이러한 방법으로 시각화(차트), OLAP 기법은 통계적인 방법을 통하여 새로운 정보를 사용자에게 제공한다[3]. 본 연구에서는 웹사이트와 메신저에 대하여 웹 사용량, 자주 가는 웹사이트 및 자주하는 메신저의 대화자에 대한 통계를 시각화 한다. 통계 결과는 사내의 네트워크 사용자 모니터링은 물론, 네트워크의 대역폭을 개별적으로 분배할 때 중요한 정보로 활용 될 수 있다. 또한 업무 이외의 다른 작업을 주로 하는 사용자에 대하여 업무 조율을 위한 자료로 활용될 수 있다.

#### 2.1.2 연관규칙(Association Rule)

연관규칙은 데이터 마이닝의 가장 대표적인 기술이다. 예로 백화점 등에서 함께 구매한 상품에 관한 연관성을 찾아내는 기술이다. 연관 규칙을 찾아주는 알고리즘 중에서 가장 많이 쓰이는 알고리즘은 Apriori 알고리즘이다[4]. 이 알고리즘은 두 가지 단계로 구성된다. 첫 번째 단계에서는 최소 지지도 설정 값에 따라 빈도수가 높은 항목의 집합들을 찾아내는 것이다. 다음 단계에서는 이들 집합들로부터 신뢰도 설정 값을 만족하는 연관 규칙을 모두 뽑아낸다. Apriori 알고리즘에서 사용하는 중요한 법칙은, 빈도수가 높은 항목의 집합에 속한 모든 부분 집합도 빈도수가 높다는 사실이다. 예를 들어 데이터의 크기가 3인 {라면, 커피, 설탕}이 최소 지지도에 의해 빈도수가 높다고 한다면 당연히 {라면, 커피}만을 봐도 빈도수가 높고, {커피, 설탕}을 봐도 빈도수가 높다. 즉 어떤 집합이 주어졌을 때 새로운 항목을 더해주면 지지도는 절대로 전보다 증가할 수 없다. <표 1>은 Apriori 알고리즘으로 $F_k$는 크기가 k인 빈도 높은 항목 집합이고 $C_k$는 크기가 k인 후보 항목 집합이다.

<표 1>의 Apriori 알고리즘은 우선 크기 한 개의 빈도수가 높은 항목들을 먼저 구하고, 그 다음에 이것들을 이용해 크기가 두 개인 빈도수가 높은 항목들의 집합을 구하는 방식으로 수행한다. 그렇기 때문에 빈도 높은 항목 집합의 크기가 k라면 대략적으로 데이터를 k번 스캔하게 된다(집합의 크기는 그 집합에 들어있는 원소 개수를 말함). 현재 크기가 k인 항목의 빈도가 높다고 할 때, 이들을 이용해 크기가 k+1인 후보 항목들의 집합들을 먼저 구한다. 예를 들면 {라면, 커피}와 {라면, 설탕}이 크기가 2인 빈도 높은 항목들 집합에 들어 있다면, 이것으로부터 {라면, 커피, 설탕}이라는 크기가 3인 후보 항목들의 집합이 만들어진다. 이 집합의 원소로 구성된 크기가 2인 모든 부분집합이 크기가 2인 빈도 높은 항목 집합들에 다 들어있는지 점검하고, 만일 하나라도 없다면 후보에서 탈락시킨다. 이런 식으로 후보들을 만든 후에는 실제 데이터를 스캔해 후보들을 계산하고, 그런 후에 지지도를 만족하는 것들만 뽑아내 크기가 3인 후보 항목 집합을 만들어 낸다. 그 다음에는 다시 이들을 이용해 크기가 4인 후보들을 만들어 내고, 더 이상 후보 집합을 만들지 못할 때까지 같은 과정을 반복한다.

본 연구에서는 네트워크 사용자가 웹사이트에 자주 접근하는 사이트의 데이터 셋(set)을 만들고 이를 기준으로 유저별 접근 사이트의 지지도를 계산하여 최종 결과로 연관성 있는 2개의 사이트를 도출한다.

〈표 1〉 Apriori 알고리즘

```
F1 = 빈도수가 높은 항목들
for(k=1; Fk-1 ≠ {}; k++) do{
 Ck = Fk-1로부터 만들어진 새로운 후보 빈도 항목집합
 for each DB의 트랙잭션 t do
  t에 있는 Ck+1의 모든 후보 빈도 항목 집합의 카운터증가
  Fk+1 = Ck+1에 있는 후보들 중에서 최소 지지도 이상 갖는 것
} 결과= F1 ∪ F2 ∪ F3 ∪... ∪ Fk
```

2.1.3 순차패턴(Sequential Pattern)

순차 패턴은 어떤 사건이 순서대로 일어난 데이터를 분석해 빈도수가 높은 형태를 찾아내는 기술이다. 예로 대형 슈퍼마켓에서 고객들의 물품 구매의 행동패턴을 데이터베이스에 저장한다고 하자. 고객 A는 처음에는 담배와 술을, 다음날에는 담배와 신문을, 그 다음날에는 음료수와 과자를 구매한다고 하면 이러한 구매 패턴을 각각 '{담배, 술}, {담배, 신문} 및 {음료수, 과자}'라는 트랜잭션으로 나타낼 수 있다. 이들 트랜잭션의 시간적 순서를 고려하면'({담배, 술},{담배, 신문},{음료수, 과자})'와 같은 트랜잭션들의 시퀀스로 나타낼 수 있다. 그 외에 다른 고객의 구매 패턴도 시퀀스들로 데이터베이스에 저장된다. 이 때, 모든 사용자 시퀀스 중 공통으로 나타내는 시퀀스를 찾는 것이 순차패턴 마이닝 기법이다.

순차패턴 알고리즘 중 GSP 알고리즘은 <표 2>와 같다[5]. <표 2>에서 $F_k$는 길이가K인 빈번한 집합이고, $C_k$는 길이가 K인 후보 패턴 집합을 말한다.

GPS 알고리즘은 크기가 1인 후보 집합부터 시작해 빈번한 집합을 찾아간다. 즉 $C_1$->$F_1$->$C_2$->$F_2$->$C_3$의 순서로 점점 길이가 긴 빈번한 패턴을 찾아 가는 것이다.

$C_1$은 $F_0$이 존재하지 않으므로, 데이터베이스에 등장하는 모든 아이템이 될 것이며 $C_{k+1}$은 $F_k$로부터 만들어 낼 수 있다. 이와 같은 방법으로 순차 패턴을 찾아간다.

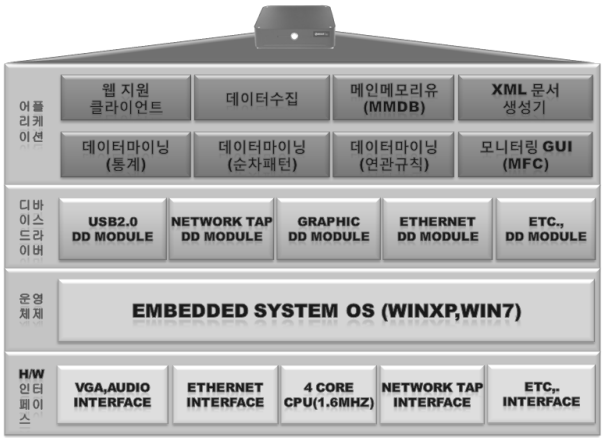
본 연구에서는 네트워크 사용자가 자주 방문한 사이트 및 사이트간의 접속 시간을 기준으로 빈번한 시퀀스 $F_k$ 집합을 만들고, 빈번한 집합 중 후보 패턴으로 $C_k$집합을 만들어 순차 패턴에 적용하여 결과 중 비율이 가장 높은 항목 3개를 도출한다.

〈표 2〉 GSP 알고리즘

$F_1$=빈번한 아이템들
for(k=2;$F_{k-1}$≠{};k++) do{
$C_k$=$F_{k-1}$로부터 만들어낸 길이 K인 후보 패턴들
for each DB에 저장된 사용자 시퀀스 C do
$C_k$ 중 C의 부분 시퀀스인 것들의 카운트를 1씩 증가
$F_k$=$C_k$중에서 카운트가 최소 지지도 이상인 것들
} 결과= $F_1 \cup F_2 \cup F_3 \cup \ldots \cup F_k$

2.2 임베디드 시스템(embedded system)

임베디드 시스템은 시스템을 동작시키는 소프트웨어를 하드웨어에 내장하여, 특수한 기능만을 가지게 한 컴퓨터 시스템이다. 임베디드 시스템의 특징으로 시스템에 키보드, 모니터, 직렬통신, 대용량 기억장치와 같은 일반적인 일부 주변장치를 지원하거나, 일부는 사용자 인터페이스를 제거해 특정 기능을 지원하지 않게 구성할 수도 있다. 그렇기에 특정 기능만을 가진 시스템 제작에 용이하고 정해진 기능만을 사용하기 때문에 최적의 성능을 발휘할 수 있게 구현 가능하다. 비용면에서도 PC용 운영체제에 비해 훨씬 저렴하다.

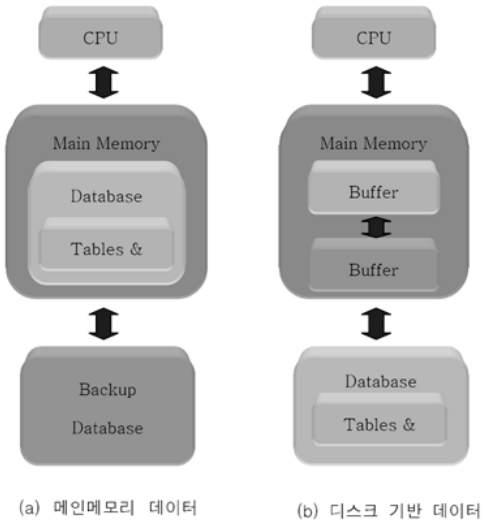



(그림 1) 행동패턴 수집용 임베디드 셋톱박스 플랫폼

(그림 1)은 행동패턴 수집용 임베디드 셋톱박스 플랫폼 하드웨어 인터페이스로 4Core CPU형태로 상용 임베디드형 제어보드로 구성하고 운영체제는 임베디드 XP를 사용한다. 임베디드 XP의 순수 OS크기는 최소 8MB로 제작할 수 있어 셋톱박스의 리소스를 최대로 어플리케이션에게 할당할 수 있다. 어플리케이션은 메모리DB, 데이터수집, 데이터마이닝, XML문서생성성기, 모니터링 및 웹 지원 클라이언트를 개발하여 포팅한다. 하드웨어 인터페이스로 네트워크-탭용 인터페이스 및 일반 통신용 이더넷 인터페이스 구성된다. 본 연구에서는 임베디드형 셋톱박스에 고속으로 패킷을 수집하고 저장하도록 고속의 메인메모리DB를 구현했다.

(그림 2)는 메인메모리DB와 디스크 기반 DB를 비교한 것이다. 이 둘의 가장 큰 차이는 메인메모리DB는 고속 처리(저장 및 액세스)를 위해 메모리에 DB가 구축되어 저장되고, 디스크기반 DB는 별도의 저장소에 데이터베이스를 저장한다는 점이다.

실시간성을 보장하기위한 데이터 쿼리(Query)를 할 경우




(그림 2) 메인메모리DB와 디스크기반DB

메인메모리DB가 월등히 우수하다[6].

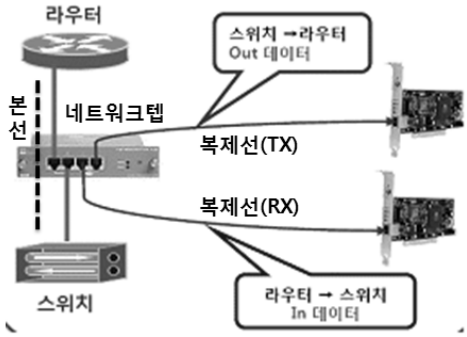
본 연구에서는 분석 및 모니터링에 최적화되고 비용측면에서도 저가로 구현 가능한 임베디드 형태의 셋톱박스로 행동 패턴 분석 시스템을 제안한다. 이 시스템은 데이터를 수집하고 데이터 마이닝은 물론 이기종간의 데이터 통신이 가능하도록 표준 XML모듈을 사용하여, 리눅스 플랫폼과 윈도우즈 플랫폼 간의 데이터를 공유한다. 구현된 시스템은 win32 환경에서 데이터 수집과 마이닝을 수행하고 리눅스 웹서버(아파치,톰캣,자바)를 통하여 원격의 관리자에게 실시간으로 행동패턴분석 결과를 전달한다. 고속의 패킷을 실시간 저장하기위해 메모리 기반의 DB를 이용한다.

### 2.3 네트워크-탭(Network-Tap)

네트워크-탭은 네트워크 본선의 데이터 흐름을 중단 없이 모니터링 할 수 있게 해주는 네트워크 패킷 복제 장비이다. 주요 특징으로 패킷의 손실 없이 전체 패킷을 모니터링 할 수 있고, 문제 처리를 위한 비정상적인 에러 패킷의 분석이 가능하며, 실시간 처리로 본선의 패킷과 비교하여 지연시간이 없다[8]. 또한 네트워크에 한번 설치하여 놓으면 아무 때나 네트워크 통신에 전혀 영향을 주지 않고 모니터링이 가능 하다. 활용분야로 IDS, PMS, IPS, DB보안, 방화벽, Performance Management Appliance 및 WAN Protocol Analysis Appliance 등이 있다.

(그림 3)은 라우터와 스위치 사이에 네트워크-탭을 설치하여 패킷의 IN/OUT(RX/ TX)으로 구분하여 복재하는 구성이다.

본 연구에서는 네트워크-탭 중 RT와 TX를 동시에 모니터링 가능한 Data Aggregation Tap를 사용하여 임베디드형 셋톱박스에서 데이터를 수집하도록 구성한다.



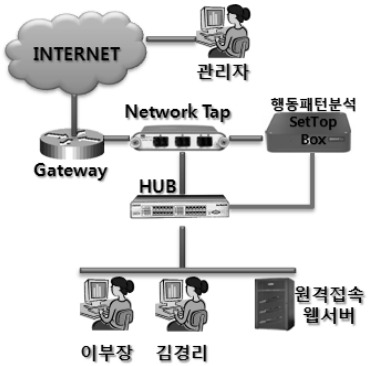

(그림 3) 네트워크-탭 구성도

## 3. 네트워크 사용자 행동패턴분석 시스템

### 3.1 시스템구성

본 연구에서 제안하는 행동패턴 분석 시스템의 구성은 (그림 4)와 같다.

네트워크는 사용자를 연결하는 허브(hub)이하의 망으로 구성된 상태에서 내부 인터넷 사용자가 웹에 접속할 때, 기




(그림 4) 분석시스템 구성도

본 망에 아무런 지장 없이, 송수신되는 패킷을 네트워크-탭을 통하여 실시간 수집이 이루어진다. 수집된 데이터는 임베디드형 셋톱박스에 구현된 고속 메인메모리 DB에 저장된다. 저장된 데이터는 데이터 마이닝을 통하여 네트워크 사용자 행동패턴을 분석하고, 결과를 로컬 또는 웹으로 관리자에게 보여준다.

### 3.2 데이터수집

데이터 수집은 네트워크 사용자의 패킷을 캡처하여 DB에 저장하는 과정으로, 네트워크 전체 데이터 전부를 수집하여 저장한다는 것은 힘든 일이다. 그래서 <표 3>과 같이 수집유형에 따른 패턴을 정의하여 수집 영역을 제한한다.

또한 고속의 패킷 수집을 위해서 메인메모리DB를 사용하여 수집하고, 수집된 결과는 디스크DB에 저장한다. <표 4>는 네트워크 패킷 수집에 사용한 메인메모리DB의 테이블 구성이다.

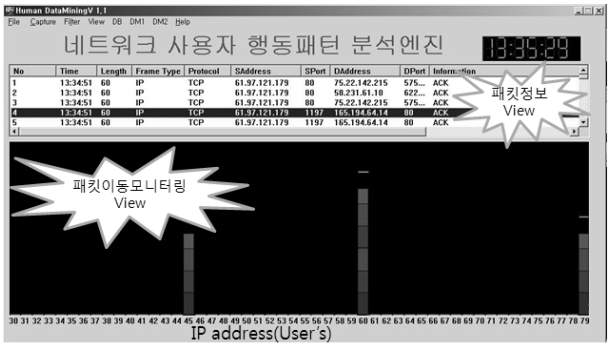
데이터 마이닝에 필요한 정보를 수집하고 데이터마이닝을 수행하기위해 행동패턴분석용 단말장치인 임베디드형 셋톱박스를 구성한다. 셋톱박스는 데이터를 수집하여 메모리DB

〈표 3〉 수집유형에 따른 패턴

| 패킷유형 | 검출 패턴 |
|---|---|
| Web용 패킷 | http://, www. com. net, co.kr, org |
| 이메일 패킷 | ADDR, mailto, to, from |
| 메신저 패킷 | CTOC, INVT |
| 기타 | 포트번호(Telnet, FTP) |

〈표 4〉 수집테이블구성

| 필드이름 | 형식 | 설명 |
|---|---|---|
| NO | int | index |
| TYPE | int | 수집유형 (1.웹, 2.이메일, 3.메신저 4.기타) |
| TIME | int | 수집시간 |
| SRC | char | 근원지IP |
| DST | char | 목적지IP |
| SIZE | int | 패킷크기 |
| DATA | char | 수집데이터(페이로드) |




(그림 5) 패킷수집 메인GUI 구성

에 데이터를 저장하고 웹서비스를 위한 XML 에이전트 포함한다.

(그림 5)는 패킷수집 상태를 표시하는 임베디드형 셋톱박스의 호스트용 GUI 어플리케이션으로 패킷정보 뷰(Packet : time, legth, src, dst, port, info)와 실시간으로 패킷의 송수신을 그래프로 보여주는 패킷 이동 모니터링 뷰로 구성된다. 모니터링 뷰의 경우 패킷의 송수신 크기가 막대 그래프의 높이에 해당되어 실시간으로 상태를 표현한다.

### 3.3 데이터마이닝

본 절에서는 관련연구에서 언급한 데이터마이닝 기법 중에 통계적 기법, 연관규칙, 순차패턴을 적용한 결과를 설명한다.

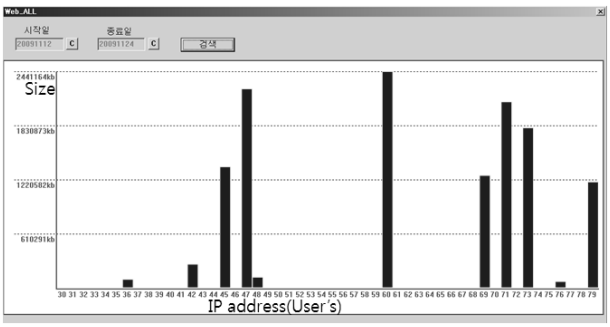
#### 3.3.1 통계적 기법

- 접속량 측정 마이닝

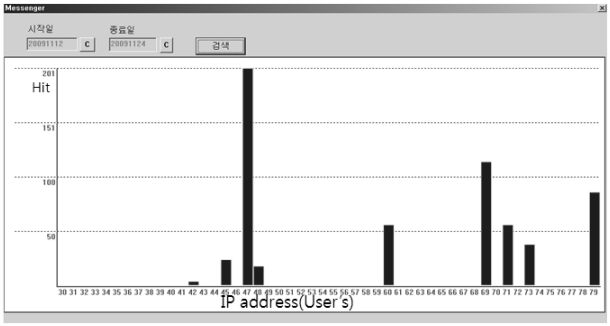
  접속량 측정 마이닝은 수집된 패킷 데이터의 분류 및 통계를 통해 작성된다. 이 통계 자료는 네트워크 사용자별 대역폭을 나타내며 향후에 개별 대역폭 할당에 기준 자료로 사용할 수 있다.

(그림 6)은 사용자(IP address)의 인터넷 접속 시 발생하는 패킷의 크기를 일정기간 동안 수집한 결과를 막대 차트로 표현한 것이다.

(그림 7)은 일정기간 동안 메신저를 사용한 유저의 접속수를 통계 차트로 표현한 것으로, 일주일 동안 누가 가장 많이 메신저를 사용하는지 확인할 수 있다. 네트워크 관리




(그림 6) Web 연결



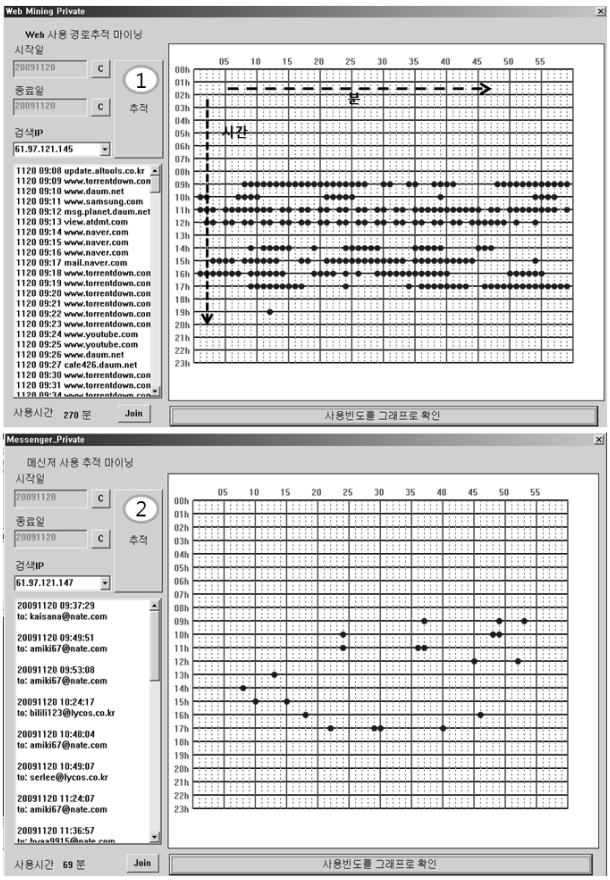

(그림 7) 메신저사용횟수

자는 사내의 패킷 데이터의 총 크기는 물론 누가 가장 많이 사용하고 적게 사용하는지를 알 수 있다. 이는 허가되지 않은 메신저 사용을 제한하기 위한 정책 자료로 활용되며 메신저를 통한 보안 사고를 예방할 수 있다. 향후에는 네트워크 대역폭을 개별로 할당 시 기준 정보로 사용할 수 있다.

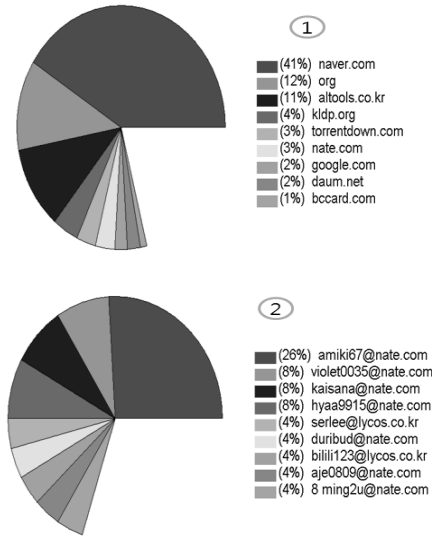
- 웹 접속 및 메신저 마이닝

  웹 접속 및 메신저 마이닝은 특정유저가 하루 동안 웹과 메신저의 사용시간 및 대상을 기록한 것이다.

(그림 8)의 ①은 유저(IP:145)의 웹에 접속한 결과를 차트




(그림 8) 웹 및 메신저 마이닝




(그림 9) 웹 사이트 분포 및 메신저 대화자분포

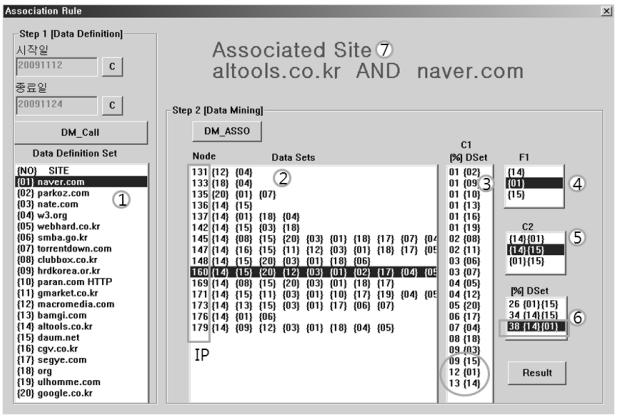
로 표현한 것으로, 사이트명과 접속시간을 24시간 기준으로 보여주고 있다. ②는 유저(IP:147)의 하루 동안 메신저에 접속한 분포를 24시간 기준으로 표현한 것이다. 전체 시간대별 분포도를 확인하여 업무에 반영할 수도 있으며 업무 이외의 시간에 나타난다면 보안상의 문제로도 점검해야 한다.

(그림 9)의 ①은 유저(IP:145)가 웹사이트에 접근하여 자주 방문한 사이트별 원형 차트로 표현한 것이고 ②는 유저(IP:147)가 메신저로 자주 대화를 나눈 대상자를 원형 차트로 표현한 것이다. 이 결과는 자주 가는 사이트와 자주 대화를 나누는 대상자의 통계 자료로 관리자에 의해 업무 연관 관계 판단 자료로 활용된다.

### 3.3.2 연관규칙

본 연구의 연관규칙은 네트워크 사용자가 웹 이용 시에 자주 방문하는 사이트 간의 연관성을 조사하는데 사용된 방법론이다. (그림 10)은 연관규칙 알고리즘을 통하여 결과가 도출되는 과정이다.

Step 1의 ①은 연관규칙을 위한 데이터 집합을 정한 것으로 접속이 빈번한 20개의 사이트를 데이터데이스에서 추출한다. Step 2의 ②는 ①에서 정의된 20개의 데이터 정의 집합을 기준으로 네트워크사용자(IP)의 웹 접근 기록을 새로운 연관 집합으로 축출한다. 이렇게 정의된 집합에서 빈도를 계산한 것이 결과 ③이다. ③에서 가장 높은 빈도 3개({14},{01},{15})를 축출한 것이 F1 집합 ④이다. ④를 가지고 구성될 수 있는 조합 ⑤만들고, ⑤를 기준으로 자주 발생하는 빈도를 ②에서 다시 조사하여 최종 결과가 ⑥이 된다. 이중 가장 빈도 높은 것이 연관성을 갖는 값으로 선정되고 ①의 SITE 정의를 바탕으로 ⑦의 결과를 도출한다. 최종 결과 이 네트워크의 대다수 웹 사용자는 'altools' 와 'naver'에 자주 연결되는 것을 확인할 수 있고 이러한 이유를 확인한 결과, 'altools' 웹 툴바 프로그램이 웹 브라우저에 설치되어 사용되고 있었고, 웹 브라우저의 초기 화면이 'naver' 이기 때문이었다.



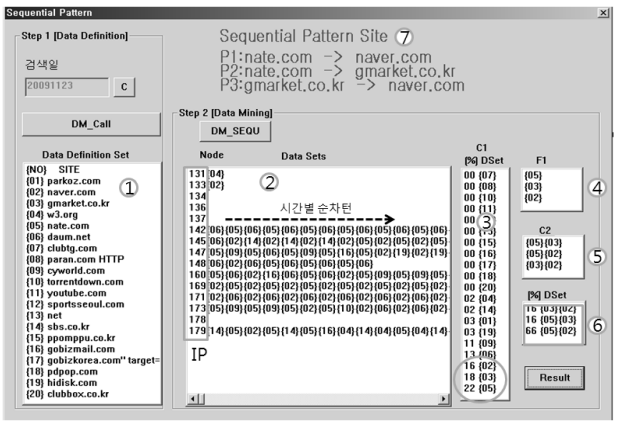

(그림 10) 연관규칙

연관규칙에 나타나는 패턴은 행동패턴에서도 비슷하게 나타나게 된다. 하지만 이것으로 패턴의 유사성과 위험성을 나누기 위해서는 정확한 업무 형태 분석과 많은 패턴 데이터 분석에 의해 정해져야하며 본 연구에서는 연관규칙의 발견을 위한 것으로 위와 같은 결과가 도출되었지만, 실제 현장 적용에는 기본 환경에서 발생하는 예외 규칙을 제외해야 한다.

### 3.3.3 순차패턴

본 연구의 순차패턴은 사내의 네트워크 사용자가 웹사이트 방문 시에 순차적으로 어떠한 사이트를 방문했는지를 확인하는 방법론이다.

(그림 11)은 순차패턴 알고리즘을 통하여 결과가 도출되는 과정이다. Step 1에선 하루 동안에 방문한 사이트를 기준으로 빈도가 가장 높은 사이트를 순서대로 ①과 같이 정의한다. 이것을 네트워크 사용자(IP)기준으로 시간별 순차패턴 알고리즘을 적용한 결과가 ②이다. 결과 ②에서 자주 발생하는 사이트를 확률로 정의한 결과가 ③이다. ③을 기준으로 빈번히 발생하는 3개의 사이트 F1(④)을 찾아서 조합한 집합이 ⑤이다. ⑤를 기준으로 ②에서 빈도를 확률로 계산한 결과가 ⑥이다. ⑥은 가장 빈도가 높은 결과 3개를 선정하여 얻은 것으로 ⑦과 같은 결과를 보여준다. 이는 P1(우선1순위)의 경우 대부분의 네트워크 사용자는 'nate'




(그림 11) 순차패턴

방문하고 이어 'naver'에 접속함을 보여준다. 이러한 결과는 이 네트워크 사용자 대부분이 메신저 'nate'와 인터넷 뷰어 초기화면이 'naver'로 되어있기 때문이다. 그 외에 웹 쇼핑 사이트인 'gmarket' 방문함을 보여주고 있다.

본 결과로 'nate'에 방문한 사용자는 곧이어 'naver'에 접속할 것이라는 예측을 할 수 있다. 이 예측을 기준으로 'nate' 접속 이후 몇 개의 사이트를 거쳐 'naver'에 접속한다거나 일정 시간이 지나도 새로운 형태의 접속 경로만을 발생하고 있다면 이상(異常)으로 판별하게 된다. 이런 결과는 구현된 시스템을 적용한 네트워크의 사용자 환경 대다수가 'nate' 메신저를 사용하고 인터넷 뷰어의 초기화면을 'naver'로 고정하고 사용하기 때문이었다. 하지만 위 연관규칙에서도 언급한 것과 같이 이것으로 패턴의 예측과 위험성을 나타내기 위해서는 정확한 업무 형태 분석과 패턴 데이터 분석에 의해 정해져야하며, 본 연구에서는 행동 패턴의 발견을 위한 것으로 위와 같은 결과가 도출되었지만, 실제 현장에 적용 시에는 이런 기본적 접속에 대한 예외처리를 고려해야 하며, 예외 처리 대상에 대한 데이터 선정에도 일정 기간과 업무 특성을 고려하여 선정하여야 한다.

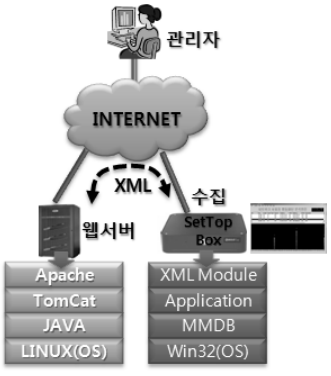
### 3.4 이기종간 데이터통신

- 시스템구성 및 XML

  데이터마이닝을 통하여 수집된 행동패턴 기록은 기본적으로는 수집전용 셋톱박스에서 확인 가능하고, 원격에서도 관리자가 웹을 통하여 실시간 확인 가능하도록 한다. 하지만 웹 서버와 연동을 위해서는 표준화된 도구가 필요하다.

(그림 12)는 마이닝 데이터를 수집하고 연산하는 셋톱박스와 원격에서 수집된 정보를 실시간 확인 가능하도록 하는 웹서버 구성도이다. 이 구성의 표준XML 적용은 다른 운영체제와의 데이터 공유를 고려한 것이다. 또한, 웹을 통한 직관적인 파악을 위해 웹-차트 사용한다.

<표 5>는 이기종간의 통신을 위한 표준xml문서의 일부분이다. XML 문서 변환은 XML 모듈을 통하여 생성하고 웹서버와 셋톱박스간의 서버 클라이언트 통신으로 XML 문서를 교환 한다.




(그림 12) 이기종간의 연결 구성도

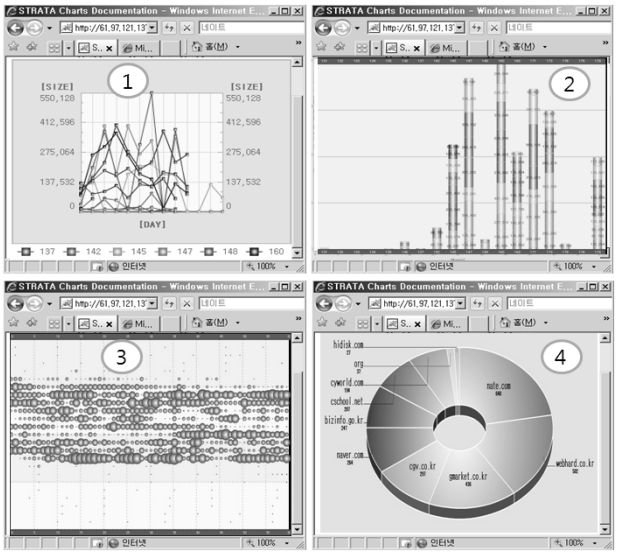
〈표 5〉 데이터통신XML

| 종류 | XML Code |
|---|---|
| 모든IP별 웹접근 | <?xml version="1.0" encoding="utf-8" ?><br>- <stratadata title="**WEB_ALL(20091112 - 20091124)**" nam<br>- <dataset><br>- <group name="**131**" color="**5c73b8**"><br><data name="**131**" color="**5c73b8**">**2141134**</data><br></group><br>- <group name="**132**" color="**5c73b8**"><br><data name="**132**" color="**5c73b8**">**5276**</data><br></group> |
| 모든IP별 MSG 기록 | <?xml version="1.0" encoding="utf-8" ?><br>- <stratadata title="**Mesenger View(20091112 - 20091124) 61.97.121.147**"><br>- <dataset><br><data name="**amiki67@nate.com**" color="**5C73B8**">**19**</data><br><data name="**cuyj66@nate.com**" color="**50A17E**">**18**</data><br><data name="**hyaa9915@nate.com**" color="**77A500**">**18**</data><br><data name="**aje0809@nate.com**" color="**FF8400**">**12**</data><br><data name="**kaisana@nate.com**" color="**DD0053**">**12**</data> |
| IP별 웹접근 | <?xml version="1.0" encoding="utf-8" ?><br>- <stratadata title="**WEB_Private(20091112 - 20091124) 61.97.1**<br>ymax="**24**" ygap="**6**"><br>- <legendset><br><data name="**HIT**" color="**5C73B8**" icon="**ff0000**" /><br></legendset><br>- <dataset><br><data name="" color="**FF0000**" icon="**circle**" x="**29**" y="**14**" /><br><data name="" color="**FF0000**" icon="**circle**" x="**30**" y="**14**" /><br><data name="" color="**FF0000**" icon="**circle**" x="**31**" y="**14**" /><br><data name="" color="**FF0000**" icon="**circle**" x="**32**" y="**14**" /> |

- 웹 서비스 동작

  (그림 13)은 행동패턴 분석 결과를 웹을 통하여 서비스한 결과이다.

(그림 13)에서 ①은 IP별(user's) 패킷사용량의 변화를 일별로 표시한다. ②는 일정기간 동안 사용자별 누적 표시를 보여준다. ③은 일주일 동안 동일한 시간대에 접근한 결과를 원의 크기의 변화로 표현한 것으로, 원의 크기가 크면 동일한 시간대에 더 자주 접근한 것이다. ④는 웹 사용자가 자주 가는 사이트를 각각 상대 비율별로 표현한 것이다. 관리 대상에 따라 사용량의 추이나 상대 사용량이나 접속 사이트의 비율 등 조건에 맞는 결과 조회가 가능하다.




(그림 13) 웹-사용자 접근서비스

## 4. 결 론

대부분의 네트워크 침입탐지를 위한 데이터마이닝 기법이 적용된 보안 알고리즘은 패킷 속성을 검사하는 방식이 주를 이루고 있다[9, 10]. 침입탐지·차단시스템과 네트워크 분석 엔진 및 보안 솔루션 대부분도 네트워크의 패킷의 헤더를 분석하여 보안정책을 세우고 정보를 보호한다. 뿐만 아니라 실제 판매되는 침입탐지·차단시스템 역시 외부로부터의 침입과 이상접근에 대한 제어가 대부분이고 내부의 네트워크를 점검하는 것은 이상패킷 발생 등을 고려한 일부에 지나지 않다. 웹 마이닝 기법 등의 로그를 분석 하는 것[11]이나 네트워크로 전달되는 명령어들을 통한 불법 침입의 검사 역시 좋은 보안 솔루션으로 활용될 수 있으나 본 연구에서 제안하는 사용자의 패턴을 탐지할 수는 없다[12]. 더욱이 내부의 공인된 사용자에 의한 접근에 대해서는 보안 정책을 수립하기 힘들다. 일부에서는 내부 보안을 위해 허가된 PC에서만 파일을 읽고 쓸 수 있게 하거나, 네트워크 자체를 외부망과 분리해서 사용하기도 하지만, 업무적 활동과 보안위반 활동을 구분하기란 쉽지 않다.

본 연구에서는 내부 사용자의 행동패턴을 통하여 바이러스 백신이나 보안장비로는 찾을 수 없는 형태의 보안 솔루션을 제안했다. 네트워크 사용자 행동패턴은 웹 접근 빈도 및 접근 시간을 데이터 마이닝 기법과 메인메모리DB를 이용하여 구현할 수 있음을 보였다. 이 결과는 내부 사용자가 본인의 주 업무를 벗어난 다른 작업을 수행하는 정도를 확인할 기준 데이터로도 활용된다. 또한 개인별 웹 사용은 물론, 메신저와 같은 비 웹 접근에서도 패킷을 분석하여 대당 네트워크에 적용할 바이러스 백신이나 보안 장비의 성능 기준을 제시할 수 있다. 웹 접근성을 고려한 사이트의 구현, 사내 부서의 네트워크 사용자의 감시와 모니터링, 방화벽 및 보안정책 기준 자료 생성 등에도 활용할 수 있다.

지금까지 관련 행동패턴의 연구의 초점은 소프트웨어와 코드 자체의 패턴에 있었다면 본 연구는 인간 자체의 행동패턴에 맞춘, 공학에 사회공학적 요소를 가미한 것으로도 주목할 만하다. 보안장비가 바이러스 백신 및 라우팅 등의 기능과 결합된 형태를 보이는 요즘 추세라면 본 연구의 결과를 응용하거나 유사한 패턴 검사 기능을 구현한 새로운 형태의 보안장비의 출현도 기대된다.

향후 연구과제로 행동패턴이라고 규정할 범위선정에 대한 오차율과 예외 항목으로 설정할 부분에 대한 연구가 더 보완되어야 할 것이다. 대형 네트워크에 적용할 시, 추가 감시용 셋톱박스를 설치한 후 클러스터 개념을 적용한 사내의 부서별 감시 체계를 구축하도록 하는 구성도 연구가 필요하다.

## 참 고 문 헌


[1] Micahael Goebel and Le Gruenwald, "A Survey of Data Mining Software Tools",ACM SIGKDD Exploration, June, 1999, Vol.1, Issue 1.

[2] 심규석, 데이터마이닝 걷히는 안개를 바라보면서,Micro Software 연재, 2001년 5월호.

[3] Galit Shumeli외2인, 데이터마이닝, 사이텍미디어, 2009.

[4] Annupan rodtook, Stanislav Makhanov,"Apriori Data Mining on Rotationally Invariant Multiresolutianal Moments for Pattern Reconition", JCIS-2006 Proceedings, October 2006.

[5] R, Agrawal and R. Srikant."Mining Sequential Patterns". In Proc. of the 11th Conference on Data engineering, Taipei, Taiwan, March 1995.

[6] Altibase, http://www.altibase.com/product/altibase/alti_ feature.jsp

[7] W3C, http://www.w3.org/standards/xml/.

[8] DATACOM, http://www.datacomsystems.com/solutions/choosing-network-taps.asp.

[9] STERRY BRUGGER University of California, Davis "Data Mining Methods for Network Intrusion Detection" June 9, 2004.

[10] Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, Pang-Ning Tan Computer Science Department, 200 Union Street SE, 4-192, EE/CSC Building University of Minnesota, Minneapolis, MN 55455, USA.

[11] Dong-Soo Yang, TaeGun Jeon, Dong-Ho Jung, Chang-Soo Kim "A Study on the Generation Algorithm of Intrusion Detection using Association Mining Technique" Dept. of Computer Science, Dept. of Computer Science and Information, PuKyong Nat'1 University.

[12] 서종원, 조제경, 이형우. "웹 공격 탐지를 위한 고속화된 웹 로그 전처리 시스템", 정보처리학회논문집 제14권 제1호, pp.969-972, 2007.



**정 세 영**

e-mail : syjeong@human2000.co.kr

2001년 금오공과대학교 전자공학과(학사)

2005년 충주대학교 전자계산학과(공학석사)

2007년 충북대학교 전자계산학과(박사수료)

2008년~현 재 휴먼앤퓨처정보통신 선임연구원

관심분야 : 네트워크 보안, 무선 네트워크, 임베디드 시스템



**이 병 권**

e-mail : sonic747@hanmail.net

1999년 한밭대학교 전자계산학과(학사)

2002년 한남대학교 컴퓨터공학과(공학석사)

2007년 충북대학교 전자계산학과(이학박사)

2007년~2010년 휴먼앤퓨처정보통신(주) 연구소장

관심분야 : 임베디드 시스템, 멀티터치, 네트워크 보안