모바일 노드의 접속 정보 유출이 없는 무선 네트워크 트래픽에 대한 순차 패턴 마이닝

송 지 영　　김 승 우　　박 상 현*

◈ 목 차 ◈

1. 서 론
2. 관련 연구
3. 제안하는 순차 패턴 마이닝 기법
4. 성능 평가
5. 결론 및 향후 연구과제

1. 서 론

인터넷이 보급되기 시작한 이후 네트워크 사용자의 급속한 증가에 따라 네트워크에 연결된 컴퓨터의 수와 네트워크를 통해 전송되는 데이터의 양이 점점 증가하고 있다. 최근 이들 네트워크 상에서 발생하는 대용량 네트워크 트래픽 을 자동화된 원격 정보 수집을 통하여 서버에 전송하여 저장, 관리하고, 수집된 데이터를 분석함으로써 유용한 정보를 추출하려는 연구가 활발히 진행되고 있다.

[1 - 3]. 그 예로 군사적, 상업적 용도 등의 다양한 응용 분야에서 서버로 전송되는 비정상적인 데이터의 흐름을 파악함으로써 외부로부터의 침입이나 인터넷 웜의 동작을 탐지하는 연구를 들 수 있다.

(표 1) Ethereal을 통해 수집한 네트워크 트래픽의 예

Timestamp	source address	source port	destination address	destination port
13:37:11.95	180.1.1.1	36872	yonsei.ac.kr	www
13:37:11.97	yonsei.ac.kr	www	180.1.1.1	36872
13:37:22.38	180.1.1.1	36915	192.168.1.3	telnet
13:37:22.39	192.168.1.3	telnet	180.1.1.1	36915

(표 1)은 Ethereal[1]을 사용하여 얻은 네트워크 트래픽의 예를 보인다. 각 엔트리는 트래픽이 발생한 시간과 송신 및 수신지에 대한 주소, 포트 번호 등의 정보로 구성된다. 이들 네트워크 트래픽은 일반적인 데이터와 비교하였을 때 다음과 같은 특성을 지닌다.

첫째, 모든 연결 가능한 네트워크 정보가 트래픽의 대상이 되므로 매우 많은 종류의 항목을 갖는다. 둘째, 네트워크 상에서의 매우 빈번한 송수신으로 인하여 새로운 데이터가 지속적으로 생성되는 대용량 데이터가 된다. 셋째, 네트워크 상에 연결된 다수의 사이트에서 발생되는 데이터로, 데이터가 분산되어 저장되며, 각 사이트는 각 개인의 네트워크 사용에 대한 트래픽 정보를 저장하게 된다.

이러한 특징을 갖는 대용량 네트워크 트래픽의 분석을 위해서는 클러스터링이나 연관 규칙 발견과 같은 다양한 데이터 마이닝 기법을 사용할 수 있다. 하

(표 2) 네트워크 트래픽에서 발견할 수 있는 패턴의 예

패턴1:	192.168.1.254로부터의 데이터 수신 → 192.168.1.254로의 데이터 송신 → 192.168.1.254로의 데이터 송신 → 192.168.1.254로의 데이터 송신
패턴2:	amazon.com으로부터의 데이터 수신 → amazon.com으로의 데이터 송신

* 연세대학교 컴퓨터과학과 부교수

☆ 본 논문은 서울시가 시행하고 서울시립대학교 "지능형 도시 사업단 (스마트 - 유비쿼터스 - 시티 사업단)"이 주관하는 "스마트시티를 위한 지능형 도시정보 컨버전스 시스템 개발"사업(10561)에서 지원을 받았습니다.

1) http://www.ethereal.com/

지만, 네트워크 상에서 발생한 항목들 간의 의미 있는 시간적 선후 관계를 발견하기 위해서는 순차 패턴 마이닝 기법을 활용해야 한다[1,2]. (표 2)는 네트워크 트래픽으로부터 발견할 수 있는 순차 패턴의 예를 보여준다. 여기서, 순차 패턴 2는 다수의 사이트에서 "amazon.com"으로부터의 데이터 수신이 발생한 후에는 "amazon.com"으로의 데이터 송신이 발생함을 나타낸다.

무선 네트워크는 네트워크 연결의 매체로 전파를 사용하는 네트워크의 한 분야이다. 무선 네트워크로 연결되는 장비는 주로 이동이 가능하고 작은 무선 모바일 기기가 되며 유비쿼터스 컴퓨팅(Ubiquitous computing)에 있어서 가장 기본이 되는 기술이다. 유비쿼터스 환경과 무선 네트워크 기술의 발달에 따라 무선 네트워크에 연결되는 기기의 수가 증가하고 있으며, 무선 네트워크를 통한 통신량 또한 증가하고 있다. 무선 네트워크로 연결된 기기에서 발생되는 무선 네트워크의 트래픽의 형태와 특징은 이미 설명한 일반적인 네트워크의 트래픽과 동일하다. 무선 트래픽의 분석은 특징적인 패턴을 발견함으로써 트래픽 발생을 예측하고, 이를 통해 요구되는 데이터를 선반입(Pre-fetching)하거나 모바일 노드의 동작을 예측할 수 있다는 장점을 지닌다. 특히 유선 네트워크에서와 마찬가지로 순차 패턴 마이닝 기법을 통해 많은 양의 무선 트래픽으로부터 시간 관계를 지니는 유용한 규칙과 패턴을 발견할 수 있다.

그러나 네트워크 트래픽에는 언제 어느 곳에 접속했다는, 사용자의 인터넷 사용 유형을 알려주는 데이터가 포함되어 있으며, 특히 무선 네트워크의 트래픽을 대상으로 수행되는 마이닝 작업은 모바일 노드 사용자의 프라이버시를 직접적으로 침해한다는 큰 문제점이 존재한다. 따라서 무선 트래픽에 대해서 마이닝을 수행하기 위해서는, 모바일 노드로부터의 데이터의 수집 과정에서 데이터의 출처를 은닉하거나 데이터를 변형하는 등의 추가적인 프라이버시 보호 기술이 요구된다. 또한 은닉 혹은 변형된 형태로 수집, 저장된 데이터를 대상으로 마이닝 결과의 정확성을 보장할 수 있는 신뢰성 있는 마이닝 기법의 개발이 요구된다.

프라이버시를 보호할 수 있는 마이닝 기법에 대해서 최근 많은 연구가 진행되고 있으나, 이들 대부분은 적은 항목을 가지는 데이터를 대상으로 하거나 소수의 사이트만을 대상으로 마이닝을 수행하는 방식이다. 따라서 무선 네트워크 트래픽의 많은 종류의 항목을 가지는 특성에 적합하지 않으며, 네트워크 트래픽이 발생하는 많은 사이트에 대해서 데이터를 처리하는 데는 적합하지 않다. 따라서 이러한 기존의 방식을 그대로 무선 네트워크 트래픽에 대해 적용할 경우, 결과의 부정확성 및 비실용성 등의 문제점을 초래할 수 있다.

본 논문에서는 기존 방법이 가지는 문제점을 해결하여 모바일 노드에 대한 프라이버시를 보호할 수 있는 순차 패턴 마이닝 기법에 대해 논의한다. 즉, 대용량 무선 트래픽을 대상으로 사이트에 해당하는 모바일 노드의 프라이버시를 보호하면서 마이닝 결과의 정확성과 실용성을 보장할 수 있는 효율적인 순차 패턴 마이닝 기법을 제안한다. 제안된 기법에서는 출처를 감추면서도 하나의 마이닝 서버와 같이 동작할 수 있는 N-저장소(Repository) 서버 모델을 사용한다. 이 모델은 데이터의 출처를 감춘 상태에서 빈번하게 발생한 네트워크 트래픽(즉, 빈번 항목)을 발견할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 기존 연구를 살펴본다. 3장에서는 본 논문에서 해결하고자 하는 문제를 정의하고 제안하는 기법에 대하여 설명한다. 4장에서는 네트워크 상에서 발생한 실제 트래픽을 대상으로 다양한 실험을 수행함으로써 제안된 기법의 정확성과 효율성을 검증한다. 마지막으로 5장에서는 본 논문의 공헌을 요약하고 향후 연구 과제를 간략히 기술한다.

2. 관련 연구

순차 패턴(Sequential pattern) 마이닝은 시간의 흐름에 따라 발생한 트랜잭션의 시퀀스들로부터 빈번하게 발생한 패턴을 찾기 위한 기법으로 제안되었다[4]. Srikant 등은 패턴을 구성하는 인접한 항목 간의 시간 간격 범위가 주어지고 항목 간에 종속 관계가 존재하는 일반화된 순차 패턴 문제를 정의하였으며, 문제를 해결하기 위한 GSP 알고리즘을 제안하였다[5].

Lee 등은 순차 패턴 마이닝 기법을 이용하여 침입

당한 상태와 정상 상태의 네트워크 트래픽을 비교 분석함으로써 침입 시에만 발생하는 패턴을 추출한 후, 이를 기반으로 침입 탐지 모델을 제안하여 네트워크 트래픽에 대한 마이닝 결과의 유용성을 보였다[1]. 이후에도 네트워크 트래픽을 마이닝을 통해 분석하려는 연구는 지속적으로 수행되었다[2,3].

최근에는 무선 트래픽을 분석하려는 연구들이 진행되었다. Teymori 등은 무선 네트워크를 사용하여 인터넷을 사용할 때의 트래픽을 분석하여 실제 네트워크 플로우(Flow) 분포의 특성을 알아내고자 하였다[6]. Liang의 연구는 무선 애드 혹 네트워크(Ad - hoc network) 상에서 발생되는 트래픽을 대상으로 퍼지 논리(Fuzzy logic)를 사용해 앞으로 발생한 트래픽을 추론하는 방법을 제안하였다[7].

무선 트래픽의 분석에는 데이터 마이닝을 사용할 수 있으나 기존의 네트워크 트래픽을 마이닝하여 분석하는 연구들은 네트워크 트래픽을 모두 한 곳으로 모아야한다. 때문에, 무선 네트워크에 직접 적용하는 경우 마이닝 과정에서 모바일 노드의 사용자 개개인의 인터넷 사용 유형과 같은 정보가 노출되어 모바일 노드 사용자의 프라이버시가 침해될 수 있다는 문제점을 갖는다.

Clifton 등은 마이닝 과정에서 프라이버시의 침해가 발생할 수 있다는 문제를 지적하였으며[8], 이후 이러한 문제점을 해결하기 위하여 프라이버시를 보장하면서 마이닝을 수행하기 위한 많은 연구가 진행되고 있다[9 - 12]. 이러한 연구는 크게 두 가지로 분류될 수 있다.

첫 번째는 데이터를 수집하는 과정에서 데이터를 변형하여 프라이버시를 침해하지 못하도록 제약하고, 변형된 데이터로부터 원래의 분포를 재구성하여 마이닝하는 연구이다[9 - 11]. 이 분류에는 수치 데이터에 대해 확률 분포로부터 선택된 임의의 값을 더한 변형된 값을 수집하고, 이 값들에 대한 확률 분포를 사용하여 실제 분포를 구하여 의사 결정 트리 분류자(Decision tree classifier)를 마이닝하는 방법이 있다[9]. 데이터 대체 기법(Retention replacement)을 통한 데이터의 변형과 재구성은 0과 1로 양분될 수 있는 데이터에 적용될 수 있는 기법으로, 데이터 수집 과정에서 p의 확률로 원래의 데이터를, (1 - p)의 확률로 변형된 데이터를 수집한다. 수집한 데이터로부터 0과 1에 대한 발생 빈도를 집계한 후 집계된 빈도와 확률 p를 가지고 실제 0과 1의 분포 정도를 계산할 수 있다[10]. 그러나 이 기법들은 데이터의 값이 수치 데이터나, 두 가지 값을 가지는 데이터에만 적용될 수 있다는 한계가 있다. 두 가지 방법을 다양한 데이터에 대해 적용하기 위하여 연구가 진행되었으나[11,13], 데이터가 가지는 값의 종류가 커질수록 결과의 정확성이 감소한다는 문제점이 남아있다.

두 번째는 여러 집단의 데이터를 마이닝하기 위하여 각각의 사이트가 마이닝 과정에 참여하여 프라이버시를 침해할 수 있는 데이터는 사이트가 직접 처리하고, 서버가 프라이버시를 침해하지 않는 중간 과정의 결과를 모아 최종 결과를 얻는 기법에 관한 연구이다[12]. 이 중 Kantarcioglu 등의 연구는 여러 사이트에서 각 집단이 지니는 개개의 데이터에 대한 프라이버시를 보호하면서 연관 규칙을 찾는 기법으로, 상호 암호화(Commutative encryption) 방식을 사용하여 데이터를 수집한 후, 시큐어 섬(Secure sum) 방식을 사용하여 각 사이트에서의 데이터 발생 빈도를 구함으로써 연관 규칙을 발견한다[12]. 그러나 이 방식에서 사용되는 두 가지의 연산은 전체 사이트를 연결하는 사이클을 따라 순차적으로 데이터를 전송해야 하므로 사이트가 많은 경우에는 비효율적이라는 한계점이 있다.

요약을 하면, 기존의 연구들을 대용량의 무선 네트워크 트래픽에 적용할 경우 다음과 같은 문제점이 존재한다. 첫째, 무선 트래픽은 데이터의 종류가 다양하므로 기존의 알고리즘을 그대로 적용하기 어려우며 데이터를 원래대로 복구하는 과정에서 정확성이 감소하여 원하는 마이닝 결과를 얻을 수 없다. 둘째, 무선 네트워크 상에는 매우 많은 수의 모바일 노드가 사이트로서 존재하므로 소수만을 대상으로 하는 마이닝 기법은 실용성 면에서 한계를 갖는다.

3. 제안하는 순차 패턴 마이닝 기법

이번 장에서는, 본 논문에서 제안하는 대용량의 무선 네트워크 트래픽을 대상으로 모바일 노드의 프라이버시를 보호하면서 마이닝 결과의 정확성, 실용성 등을 보장할 수 있는 효율적인 순차 패턴 마이닝 기법에 대해서 설명한다.

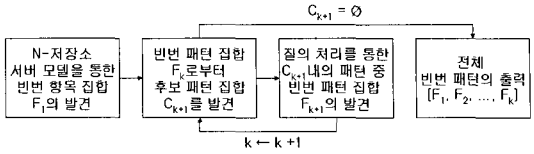
3.1 문제 정의

무선 네트워크 트래픽은 Ethereal과 같은 tcp/ip 데이터 캡쳐 프로그램을 사용하여 수집한다. Ethereal로부터 얻을 수 있는 정보는 (표 1)에서 살펴본 바와 같이 해당 트래픽이 발생한 시점의 타임스탬프와 데이터를 전송한 측과 수신한 측에 대한 주소 및 포트 번호를 포함한다. 본 논문에서는 이러한 무선 트래픽을 대상으로 각 모바일 노드의 데이터를 공개하지 않은 상태에서 (표 2)와 같은 순차 패턴을 발견하고자 한다. 이를 위하여 우선 (표 1) 형태의 무선 네트워크 트래픽을 (표 3)의 형태로 재구성한다. (표 3)에서 "out"은 모바일 노드가 데이터를 송신한 경우를, "in"은 데이터를 수신한 경우를 나타낸다. (표 1)과 (표 3)을 비교해 보면 (표 1)의 무선 네트워크 트래픽에는 포트 정보가 포함되어 있지만 (표 3)의 데이터에는 포트 정보가 포함되어 있지 않음을 알 수 있다. 포트 정보를 생략한 것은 무선 트래픽의 형태를 단순화하여 발생 가능한 무선 트래픽의 종류를 줄이기 위한 것이다. 이를 통해 빈번 패턴이 발생할 확률을 높일 수 있다.

이렇게 재구성된 무선 트래픽을 대상으로 각 트래픽 간의 의미 있는 시간적 선후 관계를 발견하기 위해서는 각각의 트래픽을 하나의 항목으로 표현한 후 순차 패턴 마이닝 기법을 적용해야 한다[1,2]. 이 때 다음과 같은 두 가지 이유로 인해 인접한 두 항목간의 최대 시간 간격을 설정하는 것이 바람직하다. 첫째, 인접한 두 항목간의 최대 시간 간격을 설정하지 않으면 고려해야 하는 순차 패턴의 수가 너무 많아진다. 둘째, 인접한 두 항목간의 시간 간격이 너무 큰 경우에는 두 항목이 연관되어 있다고 보기 어렵다. 이러한 문제점들을 고려하여 본 논문에서는 인접한 두 항목간의 시간 간격이 반드시 시스템에 의해 미리 정의된 최대 시간 간격(MaxGap)보다 작거나 같아야 한다는 제약 조건을 부여한다.

(표 3) 재구성된 네트워크 트래픽의 예

Timestamp	In/Out	address
13:37:11.95	out	amazon.com
13:37:11.97	in	amazon.com
13:37:22.38	out	192.168.1.3
13:37:22.39	in	192.168.1.3



(그림 1) 제안하는 기법의 구성

위의 내용을 정리하면 본 논문에서 해결하고자 하는 문제는 다음과 같이 정의된다. 입력으로 t개의 사이트 T1, T2, ... , Tt, 인접한 두 항목간의 최대 시간 간격 MaxGap, 최소 지지도 MinSup이 주어지면, 인접한 두 항목간의 시간 간격이 MaxGap 이하이면서 지지도가 MinSup 이상인 모든 순차 패턴을 발견한다. 이 때 각 사이트는 (표 3)의 형태로 데이터를 저장하고 있으며 데이터를 외부에 공개하지 않는다고 가정한다.

3.2 전체적인 흐름

본 논문에서 제안하는 기법은 (그림 1)에서와 같이 4단계로 구성된다. 단계 1에서는 N - 저장소 서버 모델을 사용하여 빈번 항목의 집합, 즉 길이가 1인 빈번 패턴의 집합 F1을 발견한다. 다음으로 k를 1로 설정한 후 단계 2로 진행한다. 단계 2에서는 길이가 k인 빈번 패턴의 집합 Fk를 셀프 조인하여 길이가 k+1인 후보 패턴의 집합 Ck+1을 구한다. 만약 Ck+1이 공집합이면 단계 4로 진행하고, 그렇지 않으면 단계 3으로 진행한다. 단계 3에서는 Ck+1에 속한 각 후보 패턴에 대하여 그 후보 패턴의 발생 여부를 문의하는 질의를 모든 사이트에 보낸 후 결과를 집계하여 지지도를 구한다. 다음으로, 지지도가 MinSup 이상인 후보 패턴만을 선택하여 집합 Fk+1에 포함시킨 후, k를 1만큼 증가시키고 단계 2로 되돌아간다. 단계 4에서는 F1, F2, ..., Fk를 출력하고 마이닝 작업을 종료한다.

3.3 N - 저장소 서버 모델을 통한 빈번 항목의 발견

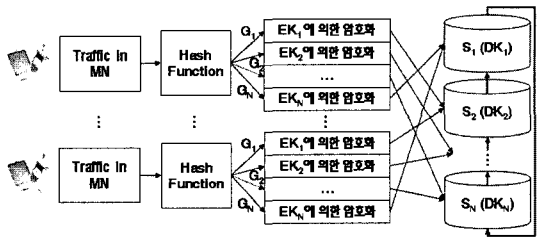
본 논문에서 제안하는 N - 저장소 서버 모델은 N개

의 서버 {S1, S2, ... , SN}와 N개의 암호화 키와 복호화 키의 쌍 {(EK1, DK1), (EK2, DK2), ... , (EKN, DKN)}으로 구성된다. 각 사이트는 N개의 암호화 키를 모두 보관하고 있으며, 서버 Si는 복호화 키 DKi만을 보관하고 있다. 빈번 항목을 발견하기 위하여 N-저장소 서버 모델은 다음과 같이 작동한다.

1) 각 사이트는 (표 3) 형태의 트래픽들을 해쉬(Hash) 함수를 사용하여 N개의 그룹 (G1, G2, ... , GN)으로 분할한다.
2) 각 사이트는 1부터 N까지의 각 i에 대하여 Gi에 속한 각각의 트래픽을 암호화 키 EKi를 사용하여 암호화 한다.
3) 각 사이트는 1부터 N-1까지의 각 i에 대하여 Gi에 속한 각각의 암호화 된 트래픽 을 서버 Si+1에 전송한다. GN에 속한 각각의 암호화 된 트래픽은 서버 S1에 전송한다. 각각의 서버 Si는 EKi로 암호화 된 트래픽만을 복호화 할 수 있으므로 각 사이트의 프라이버시는 보호된다.
4) 각각의 서버 Si는 전송받은 암호화 된 데이터로부터 같은 값을 가지는 데이터의 발생 횟수를 집계하여 빈번 항목을 생성한다.
5) 암호화 된 빈번 항목을 원래 데이터로 복호화 할 수 있도록 서버로 재전송한다. 즉, 2부터 N까지의 각 i에 대하여 서버 Si에 저장되어 있는 암호화 된 빈번 항목을 서버 Si-1로 재전송한다. 서버 S1에 저장되어 있는 암호화 된 빈번 항목은 서버 SN으로 재전송한다.
6) 각각의 서버 Si는 암호화 된 빈번 항목을 자신이 가지고 있는 복호화 키 DKi를 사용하여 복호화 한다.

3.4 길이가 2 이상인 빈번 패턴을 발견하는 알고리즘

N-저장소 서버 모델을 사용하여 빈번 항목, 즉 길이가 1인 빈번 패턴을 모두 발견한 후에는, 길이가 2 이상인 빈번 패턴을 차례로 발견해야 한다. 이를 위해 먼저 N개의 서버 중 하나를 마이닝 서버로 지정한 후, 길이가 1인 모든 빈번 패턴을 마이닝 서버로 전송한다. 마이닝 서버는 자신이 발견한 길이가 1인 빈번



(그림 2) 해쉬 함수에 따른 데이터의 분할 및 암호화 과정

패턴과 다른 서버로부터 전송받은 길이가 1인 빈번 패턴을 집합 F1에 저장하고 k를 1로 설정한 후 다음 알고리즘을 수행한다.

1) Apriori 알고리즘[4]을 적용하여, 길이가 k인 빈번 패턴의 집합 Fk를 셀프 조인하여 길이가 k+1인 후보 패턴의 집합 Ck+1을 구한다. 만약 Ck+1이 공집합이면 단계 5로 진행하고, 그렇지 않으면 단계 2로 진행한다.
2) 마이닝 서버는 Ck+1에 속한 각 후보 패턴에 대하여 그 후보 패턴의 발생 여부를 문의하는 질의를 모든 사이트에 보낸다.
3) 각 사이트는 자신이 저장하고 있는 네트워크 트래픽들로부터 후보 패턴의 발생 여부를 결정한다. 후보 패턴이 발생한 경우에는 결과 값으로 1을, 그렇지 않은 경우에는 결과 값으로 0을 마이닝 서버에 전달해야 하지만, 사이트의 프라이버시를 보호하기 위해서 기존의 데이터 대체 기법을 적용하여 변형한 값을 결과로 전달한다. 즉, 마이닝에 참여하는 모든 사이트 및 서버에 의해 그 값이 미리 약속된 p의 확률로 원래의 결과를, 1-p의 확률로 변형된 결과를 마이닝 서버에 전달한다.
4) 마이닝 서버는 0으로 응답한 사이트의 수와 1로 응답한 사이트의 수를 집계한다. 다음으로, 집계 결과와 확률 p를 이용하여 실제 0과 1의 분포 정도를 계산한 후 그 값을 최소 지지도 MinSup과 비교함으로써 각 후보 패턴의 빈번 여부를 결정한다. 최종적으로 Ck+1에 속한 후보 패턴 중에서 빈번하다고 판명된 것만을 선택하여 집합 Fk+1에 포함시킨 후, k를 1만큼 증가시키고 단계 1로 되돌아간다.

5) Ck+1이 공집합이라는 것은 더 이상 빈번 패턴을 발견할 수 없다는 것을 의미하므로 F1, F2, ... , Fk를 출력하고 마이닝 작업을 종료한다.

4. 성능 평가

본 장에서는 실험에 의한 성능 평가를 통하여 제안된 마이닝 기법의 우수성을 보인다. 4.1절에서는 실험 환경을 설명하고, 4.2절에서는 실험 결과를 분석한다.

4.1 실험 환경

본 실험에서는 무선 네트워크 트래픽과 동일한 형태와 특징을 지니는 유선·네트워크 트래픽을 사용하여 제안된 기법의 성능을 실험한다. PC에서 인터넷 사용으로 인하여 발생한 실제의 네트워크 트래픽을 패킷 캡쳐 프로그램 Ethereal을 사용하여 수집하고, 수집된 총 5,024,295개의 데이터로부터 인터넷 사용과 직접 관련 있는 tcp/udp 패킷에서 트래픽 발생 시각, 송수신 여부, 송수신지에 대한 주소만을 추출하여 736개의 IP 주소로 구성된 총 747,000개의 트래픽을 원본 데이터 및 질의 데이터로 사용한다. 평균 트래픽간의 발생 시각 간격은 461.38msec이다.

성능 평가는 다음 두 가지의 기법을 대상으로 한다.

Naive 방식은 원본 네트워크 트래픽을 직접 액세스하여 차례로 스캔하면서 빈번 패턴을 검색하는 방식으로, 제안하는 프라이버시 보장 방식에서 질의를 GSP 알고리즘[5]을 사용해서 처리한 방식이다.

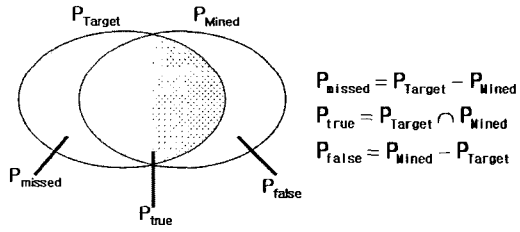
Naive 방식은 빈번 항목 집합 F1을 찾기 위해 존재하는 모든 트래픽에 대해서 데이터 대체 기법을 사용하는 방식이다. F1으로부터 2이상의 i에 대해 Fi를 찾기 위해서도 데이터 대체기법을 사용한다. N-rep방식은 N-저장소 서버 모델을 사용하여 F1을 찾는 방식이다. 2이상의 i에 대해 Fi를 찾기 위해서 데이터 대체 기법을 사용한다.

정확도 평가 지수로는 전체 정확한 빈번 패턴 중 해당 방식이 찾아내는 비율이 얼마나 되는지를 나타내는 회수율(Recall)과, 찾아낸 패턴 중 실제 정확한 패턴이 얼마나 되는지를 나타내는 정확도(Precision)를 사용한다. 트래픽에 존재하는 모든 빈번 순차 패턴의 집합을 P_{Target}이라 하고, 검색된 패턴의 집합을 P_{Mined}라고 하자. (그림 3)과 같이 P_{Target}과 P_{Mined}에 모두 포함되는 패턴의 집합을 P_{true}, P_{Target}에만 속하는 패턴의 집합을 P_{missed}, PMined에만 포함되는 패턴의 집합을 Pfalse라고는 하면, 회수율과 정확도의 정의는 다음 식과 같다.

$$\text{회수율}(Recall) = \frac{|P_{true}|}{|P_{Target}|}$$

$$\text{정확도}(\text{Precision}) = \frac{|P_{true}|}{|P_{Mined}|}$$

실험을 위한 하드웨어 플랫폼으로는 Windows XP 운영체제로 운영되고, 512MB의 메모리와 80GB (7200RPM) 디스크를 가지고 있는 Pentium IV 3.0GHz의 PC를 사용한다. 실험은 JAVA 2 Runtime Environment 1.4.2 상에서 수행한다.



(그림 3) 발견해야 하는 패턴 집합과 발견한 패턴 집합의 관계

4.2 정확도 실험 결과 및 분석

제안된 N-저장소 서버 모델의 정확도를 평가하기 위하여 해당 모델을 사용한 N-rep 방식과 기존 방식인 Naive 방식과의 마이닝 회수율과 정확도를 비교, 분석한다.

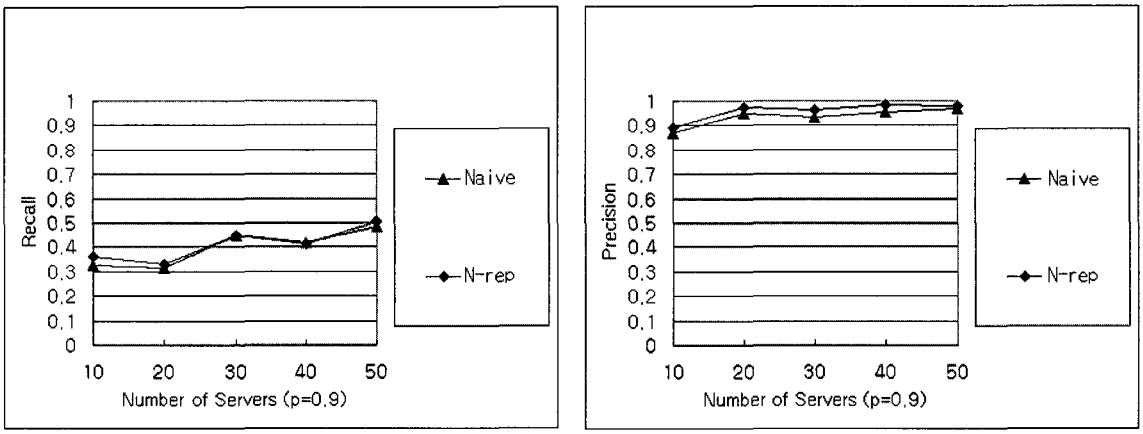
먼저, 사이트 수 변화에 따른 회수율과 정확도의 변화를 실험한다. 이 때, 변환 확률 p는 0.9로 설정한다. (그림 4)에 사이트 수를 10부터 50까지 증가시키며 회수율과 정확도를 측정한 결과를 보인다.

실험 결과에 의하면 Naive, N-rep 방식 모두 사이트 수가 증가함에 따라 회수율과 정확도가 모두 증가한다. 이는 두 방식에서 길이 2 이상의 순차 패턴을 찾기 위하여 모두 사용하는 데이터 대체 기법의 특성상 대상이 되는 사이트가 많을수록 정확도가 높아지기 때문이다. N-rep 방식의 회수율은 Naive 방식에 비해 1.01배에서 1.04배, 정확도는 1.01배에서 1.32배의 좋은 결과를 보였다. 이는 N-rep 방식에서 사용한 N-저장소 서버 모델은 빈번 항목 집합을 정확하게 찾기 때문이다.

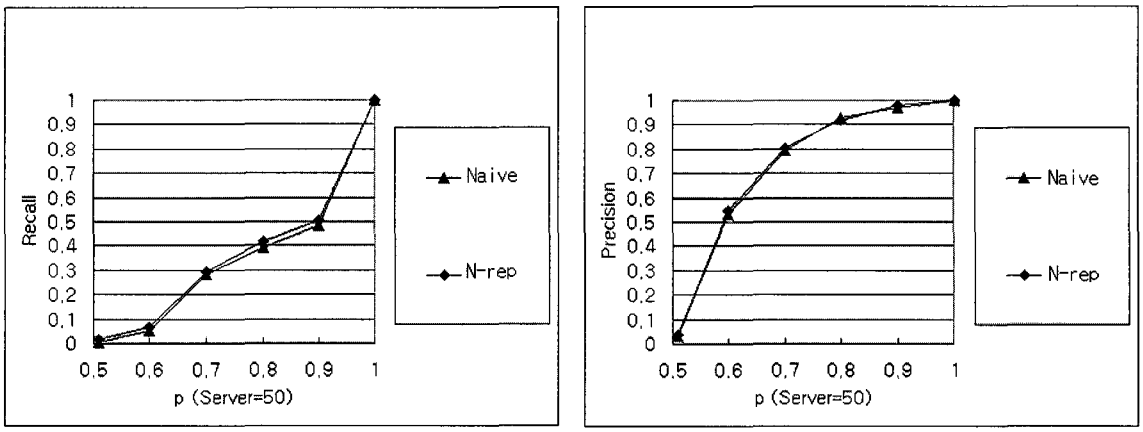
다음으로, 변환 확률 p의 변화에 따른 회수율과 정확도의 변화를 실험한다. 이 때, 사이트 수는 50으로 설정한다. (그림 5)는 p를 0.51에서 1까지 증가시키며 회수율과 정확도를 측정한 결과를 보인다. 데이터 대체 기법의 실제 분포 계산식에서 p값이 0.5인 경우 0으로 나누는 연산이 이루어지므로, p 값으로 0.5는 사용이 불가능하다[10].

실험 결과에 의하면, Naive, N-rep 방식 모두 p가 1에 가까울수록 회수율과 정확도가 모두 증가하며, 0.5에 가까울수록 감소한다. 이는 두 방식이 길이 2 이상의 순차 패턴을 찾기 위해 모두 사용하는 데이터 대체 기법의 특성에 따른 결과이다. N-rep 방식은 Naive 방식에 비해서 회수율은 1.04배에서 1.20배, 정확도는 1.01배에서 1.12배의 좋은 결과를 나타내었다.

Naive 방식은 발생할 수 있는 모든 항목을 미리 알아야 하는 방법으로 실제 무선 네트워크에서 발생한



(그림 4) 사이트 수 변화에 따른 회수율, 정확도의 변화



(그림 5) 변환 확률 p의 변화에 따른 회수율, 정확도의 변화

인터넷 트래픽에는 적용하기 힘든 방식이다. 또한, N-rep 방식이 두 가지 실험에서 모두 더 높은 정확도를 보였다.

4. 결론 및 향후 연구과제

본 논문에서는 대용량의 무선 네트워크 트래픽을 대상으로 모바일 노드의 프라이버시를 보호하면서 마이닝 결과의 정확성과 실용성을 보장할 수 있는 효율적인 순차 패턴 마이닝 기법을 제안하였다. 제안된 기법을 활용하여 여러 모바일 노드에서 빈번하게 발생하는 웹 페이지의 순차적 방문 패턴을 추출하여, 그 결과를 모바일 노드에서 웹 페이지를 선반입(Prefetch)하기 위하여 사용할 수 있다.

본 논문의 공헌을 요약하면 다음과 같다. 첫째, 무선 네트워크 상에서 발생한 무선 트래픽 에 대해 마이닝 기법을 사용하여 분석하는 방법을 제시하였다. 둘째, 하나의 마이닝 서버와 같이 동작할 수 있는 N-저장소 서버 모델과 후보 패턴의 발생 여부를 확률적으로 변형하여 전달하는 정보 유지 대체 기법을 사용함으로써 각 모바일 노드에 저장되어 있는 네트워크 데이터를 공개하지 않은 상태에서도 빈번 순차 패턴을 발견할 수 있는 방법을 제시하였다. 셋째, 무선 트래픽과 동일한 형태, 특성을 가지는 유선 네트워크 상에서 발생한 실제 트래픽을 대상으로 다양한 실험을 수행함으로써 제안된 기법의 효율성 및 정확성을 검증하였다.

향후 연구는 본 논문에서 제안한 유비쿼터스 환경의 모바일 노드 프라이버시를 보호하는 순차 패턴 마이닝 기법을 사용하여 찾아낸 순차 패턴을 활용하여, 무선 네트워크 상에서 효율적인 선반입 처리를 함으로써 무선 네트워크의 자원을 더 효율적으로 활용할 수 있는 방법을 발견하는 것이다.

참 고 문 헌

[1] W. Lee, S. Stolfo, and K. Mok, "A Data Mining Framework for Building Intrusion Detection Models," IEEE Symposium on Security and Privacy, pp. 120 - 132, 1999.

[2] S. Song, Z. Huang, H. Hu, and S. Jin, "A Sequential Pattern Mining Algorithm for Misuse Intrusion Detection," International Workshop on Information Security and Survivability for Grid (GISS2004), pp. 458 - 465, 2004.

[3] P. Dokas, L. Ertoz, V. Kumar, A. Lazarevic, J. Srivastava, and P. Tan, "Data Mining for Network Intrusion Detection," In Proceedings of NSF Workshop on Next Generation Data Mining, pp.73 - 81, 2002.

[4] R. Agrawal and R. Srikant, "Mining sequential patterns," In Proceedings of the 11th International Conference on Data Engineering, pp. 3 - 14, 1995.

[5] R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations and performance improvements," In Proceedings of the 5th International Conference on Extending Database Technology (EDBT96), pp. 3 - 17, 1996.

[6] S. Teymori and W. Zhuang, "Queue Analysis for Wireless Packet Data Traffic," In Proceedings of the 4th IFIP - TC6 Networking Conference, LNCS 3462, pp. 217 - 227, 2005.

[7] Q. Liang, "Ad Hoc Wireless Network Traffic: Self - Similarity and Forecasting," IEEE Communication Letters, Vol. 6, No. 7, pp. 297 - 299, 2002.

[8] C. Clifton and D. Marks, "Security and Privacy Implication of Data Mining," In Proceedings of the 1996 ACM Workshop on Data Mining and Knowledge Discovery, pp. 15 - 19, 1996.

[9] R. Agrawal and R. Srikant, "Privacy - preserving data mining," In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp. 439 - 450, 2000.

[10] S. Rizvi and J. Haritsa, "Maintaining Data Privacy in Association Rule Mining," In Proceedings of the 28th Conference on Very Large Data Base (VLDB'02), pp. 682 - 693, 2002.

[11] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke "Privacy Preserving Mining of

Association Rules," In Proceedings of 2002 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 217 - 118, 2002.

[12] M. Kantarcioglu and C. Clifton, "Privacy -preserving distributed mining of association rules on horizontally partitioned data," In Proceedings of the 2002 ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 24-31, 2002.

[13] R. Agrawal and R. Srikant and D. Thomas, "Privacy preserving OLAP," In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pp. 251-262, 2005.

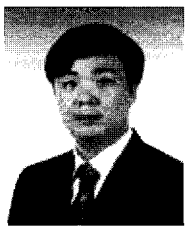
◐ 저 자 소 개 ◑

송 지 영

2005년 2월 연세대학교 컴퓨터과학과 졸업 (학사).
2005년 3월~현재 연세대학교 컴퓨터과학과 석사과정.
관심분야 : 유비쿼터스 컴퓨팅, 스트림 데이터 마이닝, 멀티미디어 데이터베이스 등
E - mail : jysong@cs.yonsei.ac.kr

김 승 우

2005년 2월 연세대학교 컴퓨터과학과 졸업 (학사).
2005년 3월~현재 연세대학교 컴퓨터과학과 석사과정.
관심분야 : 데이터 마이닝, 데이터베이스 보안, 유비쿼터스 컴퓨팅 등
E - mail : kimsw@cs.yonsei.ac.kr



박 상 현

1989년 2월 서울대학교 컴퓨터공학과 졸업 (학사).
1991년 2월 서울대학교 컴퓨터공학과 졸업 (석사).
2001년 2월 UCLA대학교 전산학과 졸업 (박사).
1991년 3월~1996년 8월 대우통신 연구원.
2001년 2월~2002년 6월 IBM T. J Watson Research Center Post - Doctoral Fellow.
2002년 8월~2003년 8월 포항공과대학교 컴퓨터공학과 조교수.
2003년 9월~2006년 8월 연세대학교 컴퓨터과학과 조교수.
2006년 9월~현재 연세대학교 컴퓨터과학과 부교수.
관심분야 : 데이터베이스 보안, 데이터 마이닝, 바이오인포매틱스, XML 등
E - mail : sanghyun@cs.yonsei.ac.kr