데이터 마이닝을 통한 네트워크 이벤트 감사 모듈 개발*

한석재** · 소우영**

요 약

최근 새로운 공격기법에 대한 대응방법의 하나로 네트워크 상황 즉, 네트워크 사용량을 분석을 통한 외부 공격 예방기법이 연구되고 있다. 이를 위한 네트워크 분석을 데이터 마이닝 기법을 통하여 네트워크 이벤트에 대한 연관 규칙을 주어 외부뿐만 아니라 내부 네트워크를 분석할 수 있는 기법이 제안되었다. 대표적인 데이터 마이닝 알고리즘인 Apriori 알고리즘을 이용한 네트워크 트래픽 분석은 과도한 CPU 사용시간과 메모리 요구로 인하여 효율성이 떨어진다. 본 논문에서는 이를 해결하기 위해서 새로운 연관 규칙 알고리즘을 제시하고 이를 이용하여 네트워크 이벤트 감사 모듈을 개발하였다. 새로운 알고리즘을 적용한 결과, Apriori 알고리즘을 적용한 시스템에 비해 CPU 사용시간과 메모리의 사용량에 있어 큰 향상을 보였다.

Development of Network Event Audit Module Using Data Mining*

Seakjae Han** · Wooyoung Soh**

ABSTRACT

Network event analysis gives useful information on the network status that helps protect attacks. It involves finding sets of frequently used packet information such as IP addresses and requires real-time processing by its nature. Apriori algorithm used for data mining can be applied to find frequent item sets, but is not suitable for analyzing network events on real-time due to the high usage of CPU and memory and thus low processing speed. This paper develops a network event audit module by applying association rules to network events using a new algorithm instead of Apriori algorithm. Test results show that the application of the new algorithm gives drastically low usage of both CPU and memory for network event analysis compared with existing Apriori algorithm.

Key words : Datamining, Apriori Algorithm, Event Audit Module

* 본 연구는 과학기술부 지역협력연구사업(R12-2003-004-01002-0) 지원으로 수행되었음.
** 한남대학교 컴퓨터공학과

1. 서 론

사이버 상의 정보보호를 위한 침입차단시술, 침입탐지기술 등 여러 가지 정보보호기술들은 알려진 취약점에 대한 예방과 탐지에 대해서는 좋은 결과를 보여주지만, 알려지지 않은 취약점이나 공격에 대해서는 적절한 대응이 쉽지 않은 단점이 있다[1]. 따라서 최근의 위협 관리 시스템[2]에서는 새로운 공격 유형에 대한 대응으로 네트워크 이벤트에 대한 분석을 보여줌으로써 네트워크 관리자가 대응할 수 있도록 하고 있다[3]. 네트워크 이벤트에 대한 분석은 외부뿐만 아니라 내부 네트워크에 대한 분석을 통해 외부의 새로운 공격 유형이나 내부의 네트워크 사용량도 분석할 수 있다.

본 논문에서는 일반적인 하나의 IP주소나 포트 번호 등의 증가에 따른 분석이 아니라 데이터 마이닝 기법을 도입해 서로의 연관 규칙을 주어 네트워크 사용량을 분석한다. 본 논문에서는 실시간 트래픽 분석에 부적합한 Apriori 알고리즘[4,5]을 대신하는 새로운 알고리즘을 제시하고 이를 이용하여 네트워크 이벤트 감사모듈을 개발하였다.

본 논문의 구성은 다음과 같다. 2장에서는 데이터 마이닝 기법으로 네트워크 트래픽 분석을 위한 연관 규칙 알고리즘에 대해 알아보고 3장에서는 실시간 네트워크 트래픽 분석에 적합한 새로운 연관 규칙 알고리즘을 제시한다. 4장에서는 새로운 연관 규칙 알고리즘을 적용한 네트워크 이벤트 감사 모듈을 개발하며, 마지막으로 5장에서는 결론을 맺는다.

2. 관련 연구

2.1 데이터마이닝

많은 양의 데이터에 함축적으로 들어 있는 지식이나 패턴을 찾아내는 기술 중에 하나인 데이터 마이닝은 비교적 최근에 연구가 시작되었고 관련 소프트웨어가 개발되고 있는 첨단 전산학 분야 중 하나라 할 수 있다. 1983년에 IBM Almaden 연구소의 Rakesh Agrawal 박사를 중심으로 Quest 데이터 마이닝 프로젝트가 시작된 이후로 선진국의 유수 연구소와 대학원을 중심으로 활발하게 연구가 되어왔다[6]. 데이터 마이닝은 데이터에서 숨겨진 패턴을 탐사하는 여러 가지 연구(연관규칙, 순차패턴, 분료, 군집화 등)중 연관 규칙 탐사가 가장 많은 연구가 이루어졌고, 그 결과인 알고리즘의 적용으로 새로운 패턴을 찾아내고 있다. 연관 규칙은 한 항목들의 그룹과 다른 항목들의 그룹 사이에 강한 연관성이 있음을 나타낸다[7].

본 논문에서는 연관규칙 알고리즘 중에 Apriori 알고리즘을 분석하여 새로운 알고리즘을 유도하는데 사용한다.

2.2 Apriori 알고리즘

연관 규칙 알고리즘 규칙을 찾아주는 알고리즘 중에서 가장 먼저 개발됐고, 또 가장 많이 쓰이는 알고리즘은 APriori 알고리즘이다[8,9]. 이 알고리즘은 두 가지 단계로 구성된다[10]. 우선 첫 번째 단계에서는 최소 지지도 설정값에 따라 빈도수가 높은 항목의 집합들을 찾아내고 그 다음 단계에서는 이들 집합들로부터 신뢰도 설정값에 만족하는 연관 규칙을 모두 뽑아낸다. Apriori 알고리즘에서 사용하는 중요한 법칙은 빈도수가 높은 항목의 집합의 모든 부분집합 또한 그 빈도수가 높다는 사실이다[4]. 예를 들어 데이터에 {라면, 커피, 설탕}이 최소 지지도에 의해 빈도수가 높다라고 가정할 때, 그 부분집합인 {라면, 커피}, {라면, 설탕}, {커피, 설탕} 역시 빈도수가 높다는 것이다.

2.3 연관규칙을 적용한 데이터베이스의 활용

연관 규칙 알고리즘은 수많은 데이터가 저장

되어 있는 데이터베이스로부터 서로 연관성이 높은 항목들을 연관시켜 그 결과값을 도출하는데 이용할 수 있다. 즉, 저장되어 있는 모든 항목을 그 빈도수와 연관시켜 최종적으로 연관집합을 도출하여 가장 빈번한 데이터사이의 연계성을 찾을 수 있다.

특정 데이터베이스에서 연관 규칙을 이용하여 연관집합을 도출하는 방법은 다음과 같다.

어떤 한 데이터베이스(D)에 저장되어 있는 구매 번호와 구매 항목이 다음 <표 1>과 같다고 가정하고, 각 항목에 대한 후보자(Candidate)를 C, 첫 번째 후보자 집합은 C1, 두 번째 후보자 집합은 C2 … n번째 후보자 집합은 Cn이라 하자.

〈표 1〉 데이터베이스 D

구매번호	구매 항목
1	{1, 3, 4}
2	{2, 3, 5}
3	{1, 2, 3, 5}
4	{2, 5}

〈표 2〉 후보자 C1과 빈발집합 F1

항목집합	지지도(%)
{1}	50
{2}	75
{3}	75
{4}	25
{5}	75

C1

⇩

항목집합	지지도(%)
{2}	75
{3}	75
{5}	75

F1

후보자 집합에서 빈발된 항목만으로 빈발(Frequent)집합을 만들 수 있는데, 첫 번째 빈발집합을 F1, 두 번째 빈발집합을 F2 … n번째 빈발집합을 Fn이라 하면, 데이터베이스를 D를 스캔하여 만든 C1으로부터의 빈발 집합 F1을 도출할 수 있다.

C1에서 가장 많은 지지도를 가진 항목들로 구성된 F1의 각 항목은 75%의 지지도를 가진다. F1에서 각 항목들의 전체 집합을 {2, 3, 5}라고 할 때 부분집합은 {2, 3}, {2, 5}, {3, 5}이며, 데이터베이스 D를 스캔하여 이 부분집합들을 포함하는 C2를 생성하고, C2로부터 빈발 집합 F2를 생성한다(<표 3>).

〈표 3〉 후보자 C2와 빈발집합 F2

항목집합	지지도(%)
{2, 3}	50
{2, 5}	75
{3, 5}	75

C2

⇩

항목집합	지지도(%)
{2, 5}	75

F2

최종적으로 빈발집합 F2에서 추출한 {2, 5}라는 연관 집합을 만들어 낼 수 있다.

3. 새로운 연관규칙 알고리즘을 이용한 이벤트 감사 모듈

3.1 Apriori를 이용한 이벤트 모듈

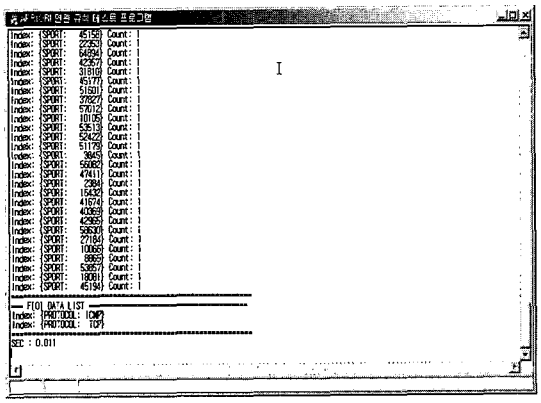
네트워크 이벤트의 입력 항목은 출발지 IP 주소, 목적지 IP 주소, 출발지 포트번호, 목적지 포트번호, 프로토콜로 하였다. 입력 항목 5개의 항목이므로 이때 나올 수 있는 연관규칙의 개수는

$\frac{5!}{2}=60$의 개수가 나올 수 있다.

네트워크 이벤트에 대한 연관 규칙 알고리즘 사용 시 지지도 대신에 이벤트의 개수로 연관된 집합을 구성하도록 하였다. 다음 화면은 Aprioiri 알고리즘으로 개발된 프로그램이다.

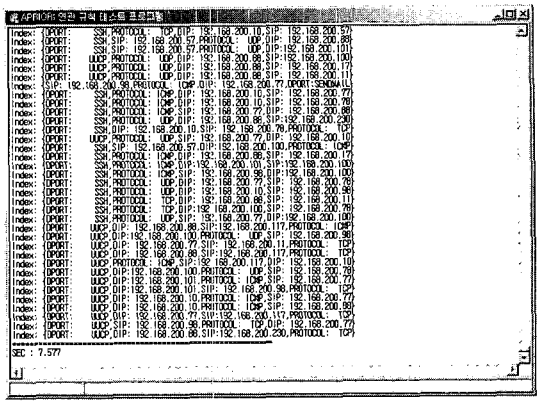
입력 데이터는 IP 주소와 목적지 포트는 빈발하게 나올 수 있도록 정해진 10가지의 데이터로 랜덤하게 나오도록 하였으며 출발지 포트는 2바이트 범위내의 양수 값 정수를 랜덤하게 나오도록 테스트하였다. 지지도의 개수는 30으로 하였다. 테스트 시 지지도의 개수를 늘리거나 줄이거나 처리속도와는 거의 무관함을 알 수 있었으므로 여기에서는 지지도의 개수를 30으로 한 테스트의 결과만을 보여줄 것이다.

(그림 1)은 Apriori 알고리즘으로 입력 데이터가 100개 일 시의 테스트 화면이다. 이때 걸린 시간은 0.011초가 걸렸으며 최종적으로 나온 빈발 집합 F[0]를 볼 수 있다.



(그림 1) 입력데이터 100개 시

(그림 2)는 Apriori 알고리즘으로 입력 데이터가 1000개 일 시의 테스트 화면이다. 이때 걸린 시간은 7.577초가 걸렸으며 100개의 입력시의 10배가 아닌 기하급수적으로 늘었음을 볼 수 있다. 메모리는 40메가 가까이 사용됨을 작업 관리자 창에서 확인하였다.



(그림 2) 데이터 입력 1000개 시

기존 Apriori 알고리즘으로 작성하였을 때 빈발집합의 개수가 n개라면 빈발집합의 부분집합으로 구성된 후보자의 집합은 $n\,C_2$로써 $\frac{n!}{(n-2)!}$의 부분집합의 개수가 된다. 만약 서로 다른 IP 주소와 포트 번호들로 구성된 빈발집합의 개수가 100개라면 부분집합의 개수는 $\frac{100!}{(100-2)!}=100\times99$가 되면 이것의 두 번째 후보자의 집합이며 두 번째 후보자의 개수가 100×99이고 두 번째 후보자의 집합이 모두 빈발집합이 된다면 세 번째 후보자의집합은 $\frac{(100\times99)!}{((100\times99)-2)!}=(100\times99)\times(100\times98)$이 된다.

위와 같은 후보자의 개수는 프로그램 연산 시 복잡도와 같으므로 프로그램이 속도 저하를 일으키게 되며 메모리의 양도 크게 된다. 실제 Apriori 알고리즘으로 작성하였을 시 10000건의 정해진 네트워크 데이터를 입력하였을 시 CPU 사용시간이 약 36초 걸렸으며 메모리 사용량은 약 530메가를 차지하였다. 실제 네트워크 이벤트를 사용했을 시 Apriori 알고리즘을 사용하면 이보다 더 많은 CPU 시간과 메모리를 요구 할 것이다.

3.2 새로운 연관 규칙 알고리즘

본 절에서는 실시간 이벤트 분석을 위한 새로운 연관 규칙 알고리즘을 제시한다. 새로운 알고리즘은 입력 시에 빈발 집합의 개수 및 후보자 집합을 계산하는 방식이다. 처음 입력 데이터가 있을 시 한 항목에 대하여 같은 항목을 찾아 개수를 증가시키고 다른 항목을 원 항목의 하부 리스트에 추가하는 방식이다.

가령 {A, B, C, D}라는 입력 데이터가 있을 시 입력 데이터를 T (Transaction)이라하고 첫 번째 입력 데이터를 T_1, 두 번째 입력 데이터를 T_2 … n 번째 입력 데이터를 T_n이라 한다. 이때 T_1의 데이터가 {A, B, C, D}라 할 때 각 항목별로 데이터 리스트를 만든다. A라는 항목에는 하부 리스트가 있으며 A 항목 리스트에는 B, C, D라는 나머지 데이터의 항목 리스트를 만든다. 또한 B의 항목에도 C, D라는 항목 리스트를 만든다. 계속하여 나머지 항목이 없을 때까지 항목 리스트를 만든다. B, C, D도 같은 방법으로 추가한다. 각 리스트의 항목에는 개수를 표현하는 변수를 둔다. 다음 표식에서 괄호안은 개수를 표현한다.

A(1) - B(1) - C(1) - D(1) - B(1) - C(1) - D(1) - C(1) - D(1) - D(1)
B(1) - C(1) - D(1) - C(1) - D(1) - D(1)
C(1) - D(1) - D(1)
D(1)

이때 리스트의 각각의 데이터 마다 개수를 셀 수 있는 정수형 변수를 사용하여 같은 항목의 데이터가 왔을 시 입력 데이터가가 입력할 때마다 개수를 증가 시키면 된다.

가령 두 번째 데이터가 {B, C, F, G}일 때, 다음과 같다.

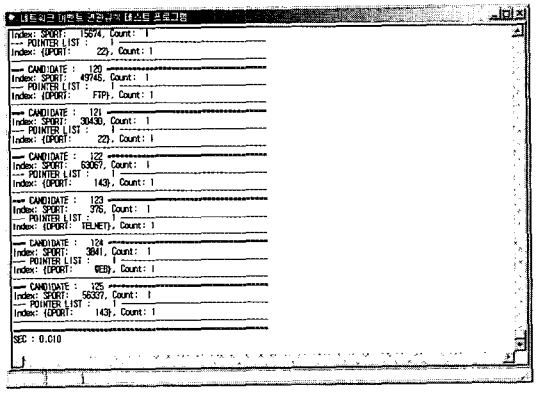
A(1) - B(1) - C(1) - D(1) - B(1) - C(1) - D(1) - C(1) - D(1) - D(1)
B(2) - C(2) - D(1) - C(2) - D(1) - D(1) - C(2) - F(1) - G(1) - F(1) - G(1) - G(1)
C(2) - D(1) - D(1) - F(1) - G(1) - G(1)
D(1)
F(1) - G(1) - G(1)
G(1)

위와 같은 결과값에서 지지도 개수를 2로 설정하였을 경우 만족하는 집합은 {B, C}이다. 위와 같은 알고리즘은 입력 데이터를 입력 시 모든 수행이 끝나게 되며 입력한 개수가 n개라면 복잡도는 $\frac{n!}{2}$로 끝나게 된다. 위에서와 같이 중복되는 연관 규칙을 막기 위해 A데이터의 입력 시 B, C, D를 입력하였으면 B 입력 시 C, D의 데이터만을 입력한다. 또한 이 알고리즘은 데이터의 순서가 없으면 하부 데이터의 구조가 깨지므로 순서를 두어 입력한다. 실험 시 네트워크 이벤트의 출발지 IP 주소, 출발지 포트번호, 목적지 IP 주소, 목적지 포트번호의 4개의 입력 항목으로 하였으므로 $\frac{4!}{2} = \frac{4\times3\times2}{2}$의 복잡도를 나타냈다. 또한 테스트 프로그램에서 Transaction 데이터는 원본을 유지할 수 있어야 하므로 데이터를 다 저장하지만 나머지 연관 규칙을 맺는 데이터 리스트에서는 해쉬 알고리즘을 사용하여 4Byte 데이터를 사용하여 메모리와 속도를 크게 줄였다. 해쉬 알고리즘을 사용하여 위와 같은 알고리즘을 사용했을 시에 네트워크 이벤트

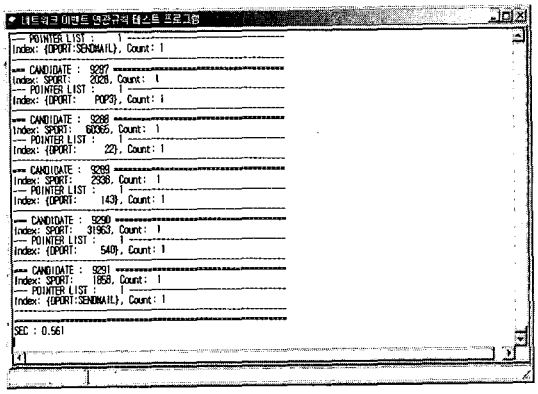
가 아닌 일반 데이터베이스의 여러 가지 데이터에 대해서도 사용할 수 있도록 하기 위해서였다.

3.3 새로운 연관 규칙 알고리즘을 적용한 이벤트 분석 테스트

다음 그림은 새로운 연관규칙 알고리즘에 의한 테스트 결과를 보여준다. 테스트시의 입력값과 지지도의 개수 등은 위에서의 Apriori 알고리즘 테스트와 동일하게 하였다. (그림 3)은 새로운 알고리즘으로 입력 데이터가 100개 일 시의 테스트 화면이다. 이때 걸린 시간은 0.010초가 걸렸음을 알 수 있다. 새로운 알고리즘으로 입력



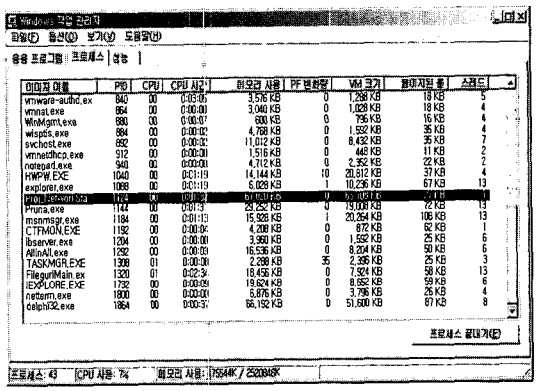
(그림 3) 입력 데이터 100개 시



(그림 4) 입력 데이터 10000개 시

데이터가 1000개 일 시의 테스트 시 걸린 시간은 0.050초가 걸렸으며 100개의 입력시의 10배에 가까운 시간이 걸렸음을 확인할 수 있다. 또한 메모리는 약 10메가 가까이 사용됨을 작업 관리자창에서 확인 할 수 있다.

위의 그림은 새로운 알고리즘으로 입력 데이터가 10000개 일 시의 테스트 화면이다. 이때 걸린 시간은 0.561초가 걸렸으며 1000개의 입력시의 10배에 가까운 시간이 걸렸음을 확인할 수 있다. 또한 메모리는 약 75메가 가까이 사용됨을 작업 관리자창에서 확인 할 수 있다.

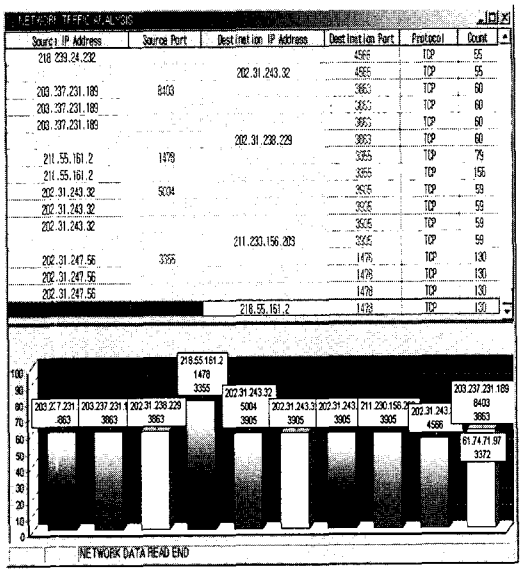


(그림 5) 입력 데이터 10000개 시 메모리 사용량

위의 같이 실험 결과로서 APriori 알고리즘을 새로운 알고리즘으로 바꿨을 시 데이터의 입력 개수가 많았을 시 현저하게 CPU 사용시간과 메모리를 줄일 수 있다.

다음 (그림 6)은 새로운 연관규칙 알고리즘을 사용하여 만든 실질적인 프로그램의 화면이다.

다음 (그림 6) 내의 출발지 IP 주소, 출발지 포트 번호, 목적지 IP 주소, 목적지 포트번호, 프로토콜, 개수등이 있으며 같은 항목으로 된 결과의 개수를 보여주고 있다. 첫 번째 결과는 218.239.24.232 출발지 IP 주소와 4556번 목적지 포트의 TCP 프로토콜로 연관된 데이터는 55개를 찾았다는 결과를 보여준다.



(그림 6) 네트워크 이벤트 연관규칙 프로그램

4. 결론 및 향후 연구방향

최근 보안상의 위협적인 네트워크 데이터를 분석하여 알려지지 않은 취약점이나 공격에 대해서는 적절한 대응을 하기 위하여 네트워크 이벤트에 대하여 분석이 필요하다. 기존의 단순히 네트워크 이벤트에 대한 개수만을 세어 분석하는 단점을 보안하고자 본 논문에서는 데이터 마이닝 기법을 사용하여 네트워크 이벤트에 대하여 분석한다. 데이터 마이닝 기법 중 Apriori 연관 규칙 알고리즘을 사용하여 분석하였으나 기존의 알고리즘이 실시간으로 분석하여야 하는 네트워크 이벤트에 대하여 과도한 CPU 사용시간과 메모리를 요구하게 되어 새로운 연관 규칙 알고리즘을 사용하여 실시간이로 네트워크 트래픽을 분석하도록 하였다. 10,000개의 이벤트를 적용한 테스트결과, Apriori 알고리즘을 이용한 트래픽분석보다 속도 면에서는 약 60배, 메모리 사용량 면에서는 약 10배 향상되었다. 향후, 이를 적용한 정보보호시스템의 연구가 요구된다.

참 고 문 헌

[1] Hyung-Jong Kim, Hong-Geun Kim, and Tae-Ho Cho, "Simulation model design of computer network for vulnerability assessment", International Workshop on Information Security Applications (WISA), Vol. 2, pp. 203-217, 2001.

[2] H. Reiser and G. Vogt, "Threat Analysis and Security Architecture of Mobile Agent based Management Systems", Proceedings of NOMS 2000 IEEE/IFIP Network Operations and Management Symposium, Honolulu, Hawaii, USA, April 2000.

[3] Myung-Sup Kim, Young J. Won, and James Won-Ki Hong, "Application-Level Traffic Monitoring and Analysis on IP Networks", ETRI Journal, Vol. 27, No. 1, pp.22-42, Feb. 2005.

[4] Maria C. FERNANDEZ, Ernestina MENASALVAS, Oscar MARBAN, "Minimal Decision Rules Based on the Apriori Algorithm", Int. J. Appl. Math. Computer Science, Vol. 11, No. 3, pp. 691-704, 2001.

[5] Ute Ziegenhain, Josef G Bauer, "Triphone Tying Techniques Combining Apriori Rules and Data Driven Methods", European Conference on Speech Communication and Technology (EUROSPEECH), Vol. 2, pp. 1417-1420, 2001.

[6] K. Leung, M. Ercegovac, and R. Muntz, "Exploiting Recongurable FPGA for Parallel Query Processing in Computation Intensive Data Mining Applications", In UC MICRO Technical Report Feb. 1999, 1999.

[7] Mara C. FERNANDEZ, Ernestina MENASALVAS, Oscar MARBAN, Jose M. PENA, Socorro MILLAN, "Minimal Decision Rules

Based on the Apriori Algorithm", Int. J. Appl. Math. Comput. Sci., Vol. 11, No. 3, pp. 691-704, 2001.

[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, pp. 914-925, December 1993.

[9] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C., May 1993.

[10] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, June 1994.

[11] Lee, Wen Ke, Stolfo, Salvatore J. and Mok, Kui W, "Algorithms for Mining, System Audit Data", Computer Science department, Columbia University.

한 석 재

1999년 단국대학교 수학과 (이학사)
2001년 한남대학교 컴퓨터공학과 (공학석사)
2005년 한남대학교 컴퓨터공학과 (박사수료)

소 우 영

1979년 중앙대학교 전자계산학과 (이학사)
1981년 서울대학교 전자계산학과 (이학석사)
1990년 미국매릴랜드대학 전자계산학과(공학박사)
1996년 한국전자통신 초빙연구원
2003년~현재 한남대학교 정보통신교육원장
1991년~현재 한남대학교 컴퓨터공학과 정교수