상이한 특성을 갖는 아이템 그룹에 대한 가중 연관 규칙 탐사

김정자* · 정희택**

Weighted Association Rule Discovery for Item Groups with Different Properties

Jung-Ja Kim* · Hee-Taek Jeong**

요 약

장바구니 분석에서, 가중 연관 규칙 탐사는 특정 상품에 대한 아이템의 중요도를 반영함으로써 더 많은 이익을 주는 정보를 규칙으로 탐사 하였다. 그러나 트랜잭션을 구성하는 아이템들이 한개 이상의 서로 다른 그룹으로 나누어진다면, 각 그룹의 특성을 반영하는 서로 다른 측정 방법으로 평가되어야 하므로 기존의 가중연관규칙 탐사 방법을 적용할 수가 없다. 본 논문에서는 이를 해결하기 위해서 가중 연관 규칙의 새로운 탐사 방법을 제안하였다. 먼저 각 아이템들은 유사한 특성에 따라 서브 그룹으로 나누고, 아이템 중요도(아이템 가중치)는 서브 그룹에 포함된 아이템들 단위로 계산한다. 이때 적용되는 여러 가중 인자들은 아이템의 특성을 반영하는 아이템 그룹별로 재 정의하였다. 제안하는 방법은 네트워크 보안 데이터에 적용하여 위험을 일으키는 요소에 대한 위험 규칙 집합을 생성함으로써 네트워크 위험관리의 정성평가와, 규칙 생성 시 적용된 가중치와 같은 여러 통계인자들에 의해서 위험도를 계산함으로써 정량평가를 가능하게 하였다. 또한 데이터 아이템들이 상이하게 구별될 수 있는 특성을 만족하는 마켓 데이터의 새로운 응용분야에 넓게 적용될 수 있다.

Abstract

In market-basket analysis, weighted association rule(WAR) discovery can mine the rules which include more beneficial information by reflecting item importance for special products. However, when items are divided into more than one group and item importance for each group must be measured by different measurement or separately, we cannot directly apply traditional weighted association rule discovery. To solve this problem, we propose a novel methodology to discovery the weighted association rule in this paper. In this methodology, the items should be first divided into sub-groups according to the properties of the items, and the item importance is defined or calculated only with the items enclosed to the sub-group. Our algorithm makes qualitative evaluation for network risk assessment possible by generating risk rule set for risk factor using network security data, and quantitative evaluation possible by calculating risk value using statistical factors such as weight applied in rule generation. And, It can be widely used for new model of more delicate analysis in market-basket database in which the data items are distinctly separated.

키워드

가중 연관 규칙 탐사, 가중 인자, 아이템 중요도, 네트워크 위험도

*전남대학교 자연과학대학 컴퓨터 정보학부 **여수대학교 인터넷 전산정보학과

접수일자 : 2004. 05. 19

I. 서 론

연관규칙 탐사는 대용량 데이터내의 데이터 아이템 집합간의 흥미 있는 연관관계나 상관성을 탐사하기 위한 방법이다. 기존의 연관규칙 탐사 방법에서, 트랜잭션 내의 모든 아이템들은 동일한 중요도를 가지고 균등하게 처리된다. 예를 들어, 의자 => 테이블(support = 0.9, confidence = 0.9) 규칙은 '의자를 사면 테이블도 같이 산다' 라는 규칙이 전체 트랜잭션 에서 지지도 90%와 신뢰도 90%로 발생하였음을 의미한다. 이때 의자와 테이블이라는 아이템은 전체 트랜잭션 데이터베이스에서 몇 번 발생하였는가의 빈발도(frequency)에 기반 하여 규칙으로서 탐사된다. 그러나, 주인의 입장에서 임의의 물건을 팔 때 되도록이면 수익성이 좋은 물건이 팔리기를 원할 것이며, 수익성이 좋은 상품이라 하면 다른 아이템에 비하여 더 중요한(important) 상품임을 뜻한다. 예를 들어, 소파와 의자라는 두 상품을 팔 때 소파의 수익성이 의자보다 더 높다면 소파 => 테이블(support = 0.6, confidence = 0.7) 규칙은 의자 => 테이블(support = 0.9, confidence = 0.9)규칙보다 지지도는 낮더라도 더 중요한 규칙이 되는 경우이다. 이와 같이 가중 연관 규칙에서는 기존의 연관규칙에 비해 각 아이템의 중요도를 반영한 규칙을 탐사하며, 판매점 데이터베이스의 경우 더 수익성을 주는 상품이 규칙으로서 탐사 될 것이다 [1][2][3][4].

본 논문에서는, 네트워크 보안관리에 가중 연관 규칙탐사를 적용하였다. 통신망 관리 측면에서 운영자는 '어느(what) 시스템이 어느(what) 요소에 얼마(support)만큼의 중요도(significance)를 가지고 이만큼(weight) 취약 한가'를 알고 싶다고 가정하자. 예를 들어, window NT 환경에서 web 서비스를 제공하는 시스템이 같은 window NT 환경에서 ftp서비스를 제공하는 시스템보다 얼마만큼 더 취약하다는 상황을 반영할 수 있어야 하는 경우이다. 기존의 연관 규칙에서는 window NT -> web 이라는 규칙과 window NT -> ftp규칙이 빈발도만으로 고려됨으로써 두 규칙이 동일한 영향력을 가지므로 위에 언급된 정성적(qualitative)인 면을 처리할 수 없다. 그러나 가중 연관 규칙에서는 각 아이템이 가중치에 의해 계량화 됨 으로써 'window NT -> web 이 window NT -> ftp보다 20% 더 취약하다'는 식의 규칙에 의하여 '해킹에 가장 취약한 서비스는 web이다' 는 사실의 추론을 할 수 있다. 이와 같이 발견된 규칙들은 네트워크 위험 관리에 있어서 보다 세부적인 접근을 가능하게 하며, 다양한 가중 인자를 이용한 네트워크 위험 수준의 수치화가 가능하다.

장바구니 분석에서 연구된 가중 연관 규칙 탐사에서, 각 아이템은 사용자에 의해 주어진 가중치를 부여 받았다 [1][2][3][4][5]. 그러나 네트워크 운영 데이터와 같이 하나의 트랙잭션을 구성하는 데이터 아이템의 성격이 상이하다면 기존의 연관규칙을 적용할 수 가 없다. 이를 해결하기 위해서는, 상이한 데이터 아이템들을 그룹지어 나누고, 각 그룹에 따르는 아이템 가중치 들이 새롭게 정의되어야 한다.

본 논문에서는 네트워크 운영데이터와 같이 서로 다른 그룹으로 구성된 데이터에 대하여, 가중 연관 규칙탐사를 적용하여 중요한 규칙을 탐사하는 새로운 방법을 제안하였다. 결과로서, 제안하는 방법론은 네트워크 보안관리 측면의 위험수준의 수치화에 의한 정량평가와 위험 규칙에 의한 정성평가를 가능하게 하였고 새로이 정의된 가중인자들에 의하여 위험도를 정의할 수 있는 네트워크 위험 분석 모델을 제시하였다.

II. 가중 연관 규칙 탐사(Weighted Association Rule Discovery)

가중 연관규칙 탐사는 기존 연관규칙 탐사와비교하여 다음과 같은 차이점을 지닌다. 첫째는 각 탐사 단계에서 다양한 가중인자(아이템 가중치, 트랜잭션 가중치)들을 사용하여 후보 아이템 집합을 생성한다. 둘째는 여러 가중 인자들에 의하여 정의된 최소 가중 지지도를 사용하여 전정된 빈발 아이템 집합을 결정 한다 [1][4][5][6].

1. 가중 연관 규칙

가중 연관 규칙은 기존의 연관 규칙에 트랜잭션을 구성하는 각 아이템에 대한 중요도/강도(importance/intensity)를 반영한 개념이다. 아이템 중요도/강도는 다양한 가중인자(weighting factor)들을 계산하는데 사용되며, 기존 연관규칙 정의에 기반 하여 가중연관규칙에 적용될 가중 인자들을 정의 할 수 있다. 다음은 다양한 가중 인자, 가중연관규칙의 특성 및 정의이다 [2][3][4]. 아이템은 마이닝의 주안점에 근거하여 상이한 가중 영역에서 가중치를 부여받는다. 가중 인자들은 허용 가능한 가중 영역(weighting space) 내에서 정의되어야 하며, 가중 영역은 가중치가 평가되는 영역이다. 본 논문에서, WSt는 내부 트랜잭션 영역(Inner-transaction space)으로서 가중아이템으로 구성된

주 트랜잭션(host transaction)를 의미한다. WSI는 아이템 영역(Item space)으로서 트랜잭션 내에 발생한 모든 아이템들의 영역이다. WST는 트랜잭션 영역(Transaction space)으로서 트랜잭션의 전 영역을 의미한다.

가중치(weights) w(i) : 아이템 집합 I={i1, i2, ..., in}가 주어지고, j={1, 2,., n} 인 곳에서 임의의 아이템 ij 의 가중치(weights) w(i)는 $0 \le w(i)$를 만족하며, 아이템집합의 중요도를 의미한다.

가중치는 아이템 집합 가중치, 트랜잭션 가중치가 있으며 아이템 집합 가중치는 아이템 가중치의 평균을 의미하며 식 (1)과 같고, 이때 | is | 는 아이템 집합의 구성 원소수를 의미한다.

$$w(is) = \frac{\sum_{k=1}^{|is|} w(i_k)}{|is|} \quad (1)$$

트랜잭션 가중치는 각 아이템집합들을 포함하는 트랜잭션들의 평균을 의미하며 식 (2)와 같다. 이때 WSt(tk)는 트랜잭션 영역 WSt 내의 k번째 트랜잭션에 대한 내부 트랜잭션 영역을 의미한다.

$$w(t_k) = \frac{\sum_{i=1}^{|WS_t(t_k)|} weight\ (item\ (i))}{|WS_t(t_k)|}. \quad (2)$$

가중연관규칙(Weighted Association Rule) : 트랜잭션의 집합을 T 라하고 T 에 속해있는 아이템의 집합 I = {i1, i2, ..., im}는 트랜잭션(T)의 부분집합으로 정의되며, 데이터 집합 X와 Y 또한 트랜잭션(T)의 집합으로 정의된다. 가중 연관규칙(WAR) X => Y 는 $X \subset I$, $Y \subset I$ 이면서, item(X) ∩item(Y) =∅을 만족한다.

빈발 아이템 집합(Large Itemset) : X => Y 로의 가중 연관 규칙에서 X 의 k-아이템 집합과 Y 의 j-아이템 집합이 주어진 최소 가중 지지도 값(minimum weighted support value) 보다 크거나 같으면 '빈발하다(large)'고 한다.

중요한 규칙(Important Rule) : X ∪ Y 가 빈발 아이템 집합이면서 주어진 최소 신뢰도 값(minimum confidence value)보다 크거나 같으면 X => Y 로의 가중 연관규칙은 중요하다(important)고 한다.

2. 가중 연관규칙 알고리즘

가중 연관 규칙 알고리즘의 탐사과정은 다음과 같다. 먼저 1단계에서 빈발 항목집합을 위한 후보 항목 집합을 생성한다. 이는 기존 연관 규칙과 동일한 과정이다. 다음 단계는 가중 연관 규칙의 정의를 만족하는 여러 가중 인자들을 계산하는 단계이다. 이 단계에서 각 후보 항목 집합들의 가중 지지도(weight support)를 계산하고 이들 가운데 주어진 최소 가중 지지도(minimum weighted support)를 만족하는 빈발 항목 집합을 결정한다. 이 과정은 항목집합의 길이에 따르는 매번의 패스에서 반복 된다 [1][3][4][5]. 본 논문에서 제안한 상이한 특성을 갖는 데이터 아이템 그룹에 대한 새로운 가중치에 대한 정의는 정의 1, 2, 3에 명시되어 있다. 그림 1은 가중 연관규칙 알고리즘을 보이고 있다.

알고리즘 : Weighted Association Rule
input : 최소 가중 지지도(minwsp), 최소 신뢰도(minconf), 트랜잭션 데이터베이스(T)

```
1. Main Algorithm (minwsp, minconf, T, minwsp, minconf)
2. L1 = {large 1-item set};
3. for (i = 2 ; Li-1≠∅ ; i++) do begin
4.  Ci = apriori-gen(Li-1); //New candidate generate//
5. forall transactions t ∈ T do begin
6. (SC, C1) = computing(T. w); //weighed factors 계산//  //SC:주어진 아이템을 포함하는 트랜잭션 수//
7. Ct = subset(Ci, t); //Candidates contained in t //
8. forall candidates c ∈ Ct do
9. c.count ++;
10. end
11. Li = {c ∈ Ci | c.count ≥ minwsp
12. end
13. Rules(SC,L) = L∪Li;
```

그림 1. 가중 연관규칙 알고리즘

III. 상이한 특성을 갖는 아이템 그룹에 대한 가중 연관 규칙 탐사

본 논문에서는 네트워크 위험 관리 문제에 가중 연관 규칙탐사를 적용하였다. 기존의 가중 연관 규칙 탐사 방법에서는 트랜잭션을 구성하는 아이템의 특성을 고려하지 않고 빈발도 만을 고려하여 최

소 지지도 값 이상의 빈발 아이템만을 선택한다. 네트워크 운영 데이터는 트랜잭션내의 각 아이템 집합이 상이한 특성을 갖는 서로 다른 데이터 아이템 그룹으로 구성된다. 이러한 상황에서는 기존의 가중 연관 규칙 탐사 방법을 적용할 수 없으므로 다음의 사항들을 고려하여야 한다. 첫째, 가중치의 정의이다. 대부분의 판매점 데이터베이스의 경우에서는 각 아이템에 대한 가중치가 초기치 로 미리 정의된다. 제안하는 방법론에서는 상이한 특성을 갖는 데이터 아이템으로 구성된 네트워크 데이터의 상황을 충분히 반영하는 타당성 있는 가중치를 정의하여야 한다. 즉 취약/위협 보고 리스트를 구성하는 각 아이템의 빈발도를 근거로 각 아이템 가중치를 재 정의하여야 하며 이는 정의 1과 정의 2에 표현되어 있다. 둘째, 후보 아이템으로부터 가중 연관 규칙을 생성하는 빈발 항목(large item)을 결정하는데 있어서 가중 인자(weighting factor)를 적용하여 규칙을 생성시킨다는 점이다. 제안하는 방법에서는 네트워크 운영데이터에 대하여 새롭게 정의한 가중 연관 규칙 탐사 방법을 통하여 주요 위험 규칙들을 발견하였고, 규칙 탐사 시 계산된 최소 가중 지지도를 위험도로 정의하였다.

1. 가중 연관 규칙을 이용한 네트워크 위험 관리 모델

표 1은 제안하는 모델의 취약/위협 데이터베이스 예이며 각 트랜잭션 아이템은 시스템(OS), 서비스, 위험수치로 구성되어있다. 트랜잭션 가중치는 정의 2에 의하여 계산된 값이다. 취약/위협 데이터베이스에서 각 트랜잭션의 아이템이 빈발하게 발생 했다는 것은 위험에 노출된 정도가 더 크다는 것으로 가정하고, 이는 가중치가 더 높음을 의미한다. 데이터베이스의 각 아이템은 가중치를 갖고 각 아이템의 가중치는 전체 트랜잭션에서 각 아이템이 발생한 빈발도와 아이템 중요도(significance)의 합으로 정의한다. 제안하는 모델에서는 표 1과 같이 각 트랜잭션의 아이템 집합 (시스템(o), 서비스(s), 중요도(r.v))내의 아이템간의 관계가, 하나의 시스템에 대해서 제공되는 서비스가 여러 개로 구성되어 있다. 그러므로 빈발도나 가중치를 계산 시 두 요소(시스템과 서비스)는 서로 다른 기준으로 정의 되어야 한다. 대부분 아이템 가중치는 도매인의 특성을 반영하여 초기치 로서 고정된 값이지만, 제안하는 모델에서는 트랜잭션을 구성하는 아이템의 성격이 서로 상이하다는 점을 반영하여 정의하여야 한다.

표 1. 취약/위협 데이터베이스의 트랜잭션 리스트 및 위험도

Transaction ID	(시스템, 서비스)	(rv)risk value	transaction wights
T1	(Window 2000, WEB, DNS)	5	1.85
T2	(Linux 7.1, FTP, DNS)	3	1.75
T3	(Window 2000, RPC, SMTP, DNS)	3	1.68
T4	(Solaris 8, WEB, DNS, FTP, Telnet)	5	1.82
T5	(Linux 7.1 WEB, DNS, FTP, Telnet, SMTP)	3	1.73
T6	(Solaris 8, WEB)	1	1.92
T7	(Windows 2000, DNS, Telnet)	3	1.76
T8	(Solaris 8, WEB, DNS, Telnet, SMTP)	3	1.79
T9	(Linux 7.1, WEB, DNS)	3	1.82
T10	(Window NT, SMTP, DNS)	3	1.59
T11	(Solaris 8, WEB, FTP, Telnet)	5	1.81
T12	(Solaris 8, WEB, Telnet, RPC)	3	1.73
T13	(Window NT, DNS, FTP, Telnet)	1	1.63
T14	(Solaris 8, WEB, FTP, SMTP)	3	1.79
T15	(Windows 2000, WEB, FTP, RPC, SMTP)	3	1,7
T16	(Linux 7.1, WEB, FTP)	5	1.8
T17	(Window NT, WEB, DNS, FTP)	3	1.69
T18	(Solaris 8, FTP, RPC, DNS)	1	1.71
T19	(Windows 2000, WEB)	3	1.9
T20	(Window NT, WEB, DNS)	5	1.72

2. 아이템 그룹에 따르는 개선된 가중 인자 (Improved Weighted Factor)

다음은 본 논문에서 제안하는 상이한 특성을 갖

는 아이템 그룹에 대한 다양한 가중인자들을 정의하고 있다.

정의 1 : 트랜잭션 데이터베이스(T)에서 시스템 집합 o(i) ={ o1, o2,....on}, $1 \le i \le n$로 구성되고, 서비스 집합 s(j) = { s1, s2,....sm}, $1 \le j \le m$으로 구성된다. 아이템 가중치(wi)는 시스템 가중치를 w(oi), 서비스 가중치를 w(si)로 표시하며 다음과 같이 정의한다. ic는 각 아이템 빈발수이며 아이템 가중치는 부그룹(sub-group)에 포함된 아이템들만으로 정의된다.

w(i) = 빈발도(i) + 중요도(rv), rv $\in$ [0..1]

$$w(Oi) = \frac{n(O_i)}{n(T(O))} + \frac{\sum rv()O_i)}{\alpha \mid ic(O_i) \mid \times nrv}$$

$$w(Sj) = \frac{n(s_j)}{n(T(S))} + \alpha\frac{\sum rv(s_j)}{\mid ic(s_j) \mid \times nrv}$$

아이템 가중치는 정의1, 2, 3에 표현되어있다. 아이템 가중치는 아이템 빈발도와 아이템 중요도의 합으로 정의된다. 아이템 그룹에 대한 빈발도는 한 아이템 그룹을 구성하는 아이템 빈발수의 합에 대한 각 아이템 빈발수로 정의된다. 아이템 중요도는 각 아이템을 포함하는 트랜잭션의 위험 수치의 합을 아이템 빈발수로 나눈값 으로 정의된다.

정의1에서 α는 아이템 가중치를 계산하는데 있어서 빈발도에 대한 중요도를 반영하기 위한 사용자 정의 변수이다. 제안하는 모델에서는 아이템 가중치를 계산하는데 있어서 α=1로 정의하였다. 이는 일반적으로 빈발도 $\ll$ risk value 의 관계가 성립함으로 이를 논리적으로 해석하면 빈발도 보다는 해당서비스나 OS의 취약/위협이 미치는 위험정도가 더 강조되는 것이 타당하기 때문이다. nrv는 중요도를 아이템 빈발도와의 균형을 위하여 0과 1사이의 값으로 정규화하기 위한 파라미터이며, 제안하는 모델에서는 rv의 최대치가 5이므로 nrv는 5로 나눈값으로 조정된다.

정의 2 : 트랜잭션 가중치(Transaction weight)는 각 트랜잭션을 구성한 아이템 집합들의 가중치 합을 의미하며 다음과 같이 정의한다.

$$w(tk) = \frac{\sum_i w(o_i)}{n \mid o_i \mid} + \frac{\sum_j w(s_j)}{n \mid s_j \mid}$$

정의 3 : 아이템 집합의 가중 지지도로서(Weighted support), 규칙 X => Y를 반영하는 트랜잭션의 집합에서 X와 Y는 $X \subset I$, $Y \subset I$ 이면서, item(X)$\cap$item(Y) =$\varnothing$을 만족한다. 이때 t는 t$\in$T인 트랜잭션을 의미하며 가중 지지도 wsp는 모든 트랜잭션 가중치의 합에 대해 후보 아이템들을 포함하는 트랜잭션의 가중치의 합으로 정의한다.

$$wsp\ (XY) = \frac{\sum_{k=1}^{|WS_T|(X \cup Y) \subseteq t_k} w(t_k)}{\sum_{k=1}^{|WS_T|} w(t_k)}$$

wsp는 전체 트랜잭션 영역(transaction space)에서 특정 아이템 집합을 포함하고 있는 내부 트랜잭션 영역에 대한 실질적인 할당(quota)을 수치화 하는 값이다. 이는 빈발 아이템을 전정(pruning)하기 위한 통계치 로서 사용된다. 또한 가중 연관 규칙에서 전정된 빈발 아이템 집합은 어쩐 빈발 아이템 집합이 빈발하면 그 부분집합도 빈발하다는 특성인 'weighted downward closure property'를 만족한다 [1][2][3]. 제안하는 방법의 정의에 의하여 표 1의 트랜잭션의 각 아이템에 대해서 빈발수(item count), 빈발도(item support), 중요도(significance), 가중치(item weight)의 계산 결과를 표 2에서 보이고 있다.

표 2. 아이템 빈발수, 빈발도, 중요도, 가중치

아이템	빈발수	빈발도	중요도	가중치
Solaris 8	7	0.35	0.6	0.95
Window 2000	5	0.25	0.68	0.93
Window NT	4	0.2	0.6	0.8
Linux 7.1	4	0.2	0.7	0.9
WEB	14	0.26	0.71	0.97
DNS	13	0.24	0.63	0.87
FTP	10	0.19	0.64	0.83
Telnet	7	0.13	0.66	0.79
RPC	4	0.07	0.5	0.57
SMTP	6	0.11	0.6	0.71

그림 2는 표 2에 계산된 각 통계인자들에 근거하여, 가중연관규칙 알고리즘에 의해 중요한 아이템 집합을 결정하는 과정을 트리로 보이고 있다. 먼저 1-아이템 항목부터 시작하여 정의1과 2에 의하여 각 아이템가중치와 트랜잭션 가중치를 구한 다음, 정의 3의 wsp 이하의 항목은 전정되고 이들의 조합에 의하여 다음단계의 빈발 항목집합을 결정해나가는 과정을 반복한다. 그림 2에서 점선으로 표현된 사각박스는 가중 지지도 이하 값으로서 전정됨을 뜻한다.

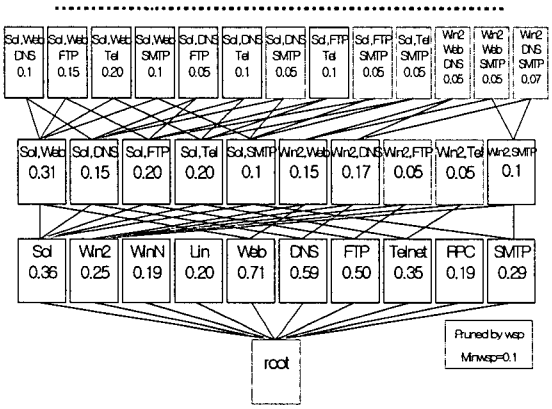

그림 2. 가중 연관 규칙 탐사에 의한 중요 아이템 집합

IV. 실험분석

제안하는 방법론은 네트워크 서비스를 제공하는 컴퓨터 시스템에 대한 취약, 위협평가에 적용하였다. 또한 제안하는 가중인자들로 재 정의된 가중 연관 규칙 알고리즘에 의하여 최소 가중 지지도이상의 빈발 아이템으로 구성된 취약/위협 규칙을 생성하였다. 이때 탐사된 규칙 집합에 정의된 wsp는 전체 트랜잭션 에서 규칙으로 생성된 중요한 아이템 집합을 정의하는 통계치 이기 때문에 이를 각 취약/위협 규칙에 대한 위험도로서 정의하였다. 결과적으로 제안하는 방법론은 컴퓨터 시스템의 보안관리측면에서 위험 수준의 질적, 양적 평가를 가능하게 하였다.

실험은 표 1의 데이터에 대하여 min_wsp를 0.5에서 1.5까지 변화시켜 실험하였다. 표 3은 min_wsp=0.1로 주어졌을 때 탐사된 22개의 위험 규칙을 보이고 있다. 예를 들어 탐사된 규칙 R4로부터 linux 7.1은 FTP 서비스에 대해 0.15만큼 위험하다, 또는 규칙 R1, R2, R3, R6, R8로부터 Solaris 8에 가장 취약한 서비스는 위험도 0.3을 갖는 WEB이라는 식의 추론을 할 수 있다. 이와 같이 제안하는 방법은 네트워크 위험 관리측면에 탐사된 규칙을 통하여 보다 의미 있는 분석을 가능하게 한다.

표 3. 가중 연관 규칙 집합의 예

번호	가중 연관 규칙	wsp
1	Solaris 8 => Telnet	0.20
2	Solaris 8 => SMTP	0.10
3	Solaris 8 => FTP	0.20
4	Linux 7.1 => FTP	0.15
5	Window NT => DNS	0.19
6	Solaris 8 => DNS	0.15
7	Linux 7.1 => DNS	0.15
8	Solaris 8 => WEB	0.30
9	Linux 7.1 => WEB	0.15
.	.	.
.	.	.
.	.	.
19	Solaris 8 => Telnet, WEB	0.20
20	Solaris 8 => Telnet, FTP	0.10
21	Solaris 8 => FTP, WEB, Telnet	0.10
22	Solaris 8 => DNS, WEB, Telnet	0.10

V. 결 론

가중 연관 규칙 탐사는 아이템의 가중치를 반영하여 규칙을 탐사하는 기존 연관 규칙 탐사 방법의 일반화된 형태이다. 기존에 마켓 데이터를 대상으로 연구된 가중 연관 규칙탐사에서는 초기에 정의된 가중치를 반영하여 중요한 규칙을 탐사한다. 그러나 트랜잭션을 구성하는 데이터 아이템의 특성이 상이할 경우 이와 같은 방법을 그대로 적용할 수 없다. 본 논문에서는 이를 해결하기 위하여 각 아이템들을 그들의 특성에 따라 여러 서브 그룹으로 나누고 각 서브 그룹의 특성에 따르는 아이템별로 가중치를 정의하는 새로운 방법론을 정의하였다.

제안하는 방법론은 가중 연관 규칙 탐사방법을

네트워크 위험 평가를 위한 새로운 응용 도메인에 적용함으로써 중요한 위험 패턴들을 규칙으로 탐사하였고, 새롭게 정의된 가중인자들을 사용하여 위험 규칙의 위험 수준을 정의하였다. 이는 보안 투자 면에서나 네트워크 운영상의 가이드라인의 제시에 대단히 효과적으로 활용할 수 있을 것이다. 또한 상이한 특성을 가진 데이터 아이템 그룹으로 구성된 마켓 데이터 응용 분야에 활용함으로써 더욱 정교한 분석을 위한 새로운 모델로 적용할 수 있다.

참고문헌

[1] Feng Tao, Fionn Murtagh, Mohsen Farid "Weighted Association Rule Mining using Weighted Support and Significance Framework", SIGKDD 2003

[2] Feng Tao, "Mining Binary Relationships from Transaction Data in Weighted Settings", PhD Thesis, School of Computer Science, Queen's University Belfast, UK, 2003

[3] W. Wang, J. Yang P. Yu, "Efficient Mining of Weighted Association Rules(WAR)", Proc. of the ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 2000, pp 270-274

[4] C.H. Cai, Ada W.C. Fu, C.H. Cheng and W.W. Kwong, "Mining Association Rules with Weighted Items" International Database Engineering and Application Symposium, 1998

[5] G.D.Ramkumar, Sanjay Ranka, and Shalom Tsur, "Weighted Association Rules : Model and Algorithm", KDD 1998.

[6] Jiawei Han and Yongjian Fu, "Discovery of Multiple-Level Association Rules from Large Databases" in the Proceedings of the 1995 Int'l Conf. on Very Large Data Bases(VLDB'95), Zurich, Switzerland, 2002, pp. 420-431

[7] N.Pasquier, Y.Bastide, R.Taouil, and L.Lakhal, "Efficient Mining of Association Rules using Closed Itemset Lattices", Information Systems, Vol. 24, No. 1, 1999, pp. 25-46.

[8] E. -H. Han, G. Karypis, and V. Kumar, "Scalable Parallel Data Mining for Association Rules", Proc. ACM SIGMOD, Tucson, U.S.A., 1997, pp. 277-288

[9] A. Savasere, E. Omiencinsky , and S. Navathe, "An efficient algorithm for mining association rules in large databases", In Proceedings of the 21st VLDB Conference, pp.432-444, Zurich, Swizerland, 1995.

[10] R. SriKant and R. Agrawal, "Mining Generalized Association Rules", In Proceedings of the 21st VLDB conference, Zurich, Swizerland, 1995.

[11] M. Klemettinen, H. Mannila, P.Ronkainen, H.Toivonen, and A.I. Verkamo, "Finding Interesting Rules from Large Sets of Discovered Association Rules", Proc. of the 3rd Intl. Conf. on Information and Knowledge Management, 1994, pp. 401-407

저자소개

김정자(Jung-Ja Kim)

1985년 전남대학교 자연과학대학 계산 통계학과(이학사)
1988년 전남대학교 자연과학대학 전산 통계학과(이학석사)
1997년~2002년 2월 전남대학교 자연과학대학 전산통계학과(이학박사)
E-mail : j2kim@chonnam.ac.kr
※ 관심분야 : 데이터베이스, 데이터 마이닝, 바이오 인포매틱스

정희택(Hee-Taek Jeong)

1992년 전남대학교 전자계산학과 졸업(이학사)
1995년 전남대학교 전자계산학과(이학석사)
1999년 전남대학교 전자계산학과(이학박사)
1999년~현재 여수대학교 인터넷전산정보전공 조교수
※관심분야 : 워크플로우 시스템, 데이터 마이닝, 분산처리 시스템