碩士學位論文

중요도를 고려한 연관규칙 탐사 알고리즘에 관한 연구

指導敎授 金 根 亨

濟州大學校 經營大學院

經營學科 經營情報 專攻

黃 炳 雄

2003年 6月

중요도를 고려한
연관규칙 탐사 알고리즘에 관한 연구

指導敎授 金 根 亨

이 論文을 經營學 碩士學位 論文으로 提出함

2003年 6月

濟州大學校 經營大學院

經營學科 經營情報 專攻

黃 炳 雄

黃炳雄의 經營學碩士學位論文을 認准함

2003年 6月

심사위원장 _____ 印

심 사 위 원 _____ 印

심 사 위 원 _____ 印

목 차

그 림 차 례

알고리즘 차례

표 차 례

Ⅰ. 서론

1. 연구 배경 및 목적

우리는 지금 정보의 홍수 속에 살고 있다. 이를 부정하는 사람은 아무도 없을 것이다. 우리의 삶 속에서 무수한 데이터를 접하고, 이를 정보화하여 의사결정을 내린다. 그러나 점점 많아지는 데이터 속에서 유용한 정보를 찾기란 점점 어려워진다. 이러한 환경 속에서 기업이 생존하고 발전하기 위해서는 지속적으로 소비자의 동향을 파악하고, 자사는 물론 경쟁사의 경영전략을 효과적으로 분석하고 대처할 수 있는 능력이 절실하게 요구된다. 이를 가능케 하는 것이 정보이다. 정보화시대의 기업경영에서 정보란 전통적인 자원들을 효과적으로 운영, 관리하며, 새로운 제품이나 서비스를 창출하는 역할을 담당하는 또 다른 자원으로서, 이것을 수집하고 활용할 수 있는 기업만이 경쟁에서 살아남고 우위를 확보하게 된다.[1)]

오늘날 데이터베이스 기술의 발전으로 이를 사용하는 업무가 급속히 늘어나면서 저장되는 데이터 양 또한 폭발적으로 증가하고 있다. 정보기술의 빠른 발전은 업무의 자동화를 촉진시켜 엄청난 양의 데이터를 전자적으로 수집하고 보관하는 기능을 가능케 했다. 데이터 수집과 저장 기술의 발달, 데이터베이스관리시스템과 데이터웨어하우스[2)] 기술의 광범위한 사용은 기업내부에 대량의 데이터를 축적할 수 있도록 하였으며, 기업들도 축적된 데이터를 의사결정에 필요한 새롭고 가치있는 정보와 지식을 획득할 수 있는 잠재적인 원천

1) 이재규, 최형림, 김현수, 이경전, "전자상거래원론", 법영사, 1999.
2) 조재희, 박성진, "데이터 웨어하우징과 OLAP", 대청, 1996.

으로 인정하고 있다. 기업들이 급변하는 경영환경에서 기업의 경쟁력을 강화하기 위해서는 축적된 데이터를 분석하고 정보와 지식을 획득 할 수 있는 능력과 정보기술을 보유해야 한다. 그러나 1990년에는 데이터를 분석하고 정보와 지식을 획득하는 능력이 데이터를 획득하고 저장하는 능력에 훨씬 미달하는 "데이터 과잉 문제 (Data Glut Problem)"가 발생하였다.[3)]

이러한 데이터 과잉 문제는 방대한 양의 데이터에 내재된 정보와 지식을 "발견" 또는 "마이닝"하는 능력의 제고에 의해서 해결 될 수 있는데, 데이터 마이닝은 이런 요구사항을 충족시키는 새로운 정보 기술의 활용방법이다. 특히 데이터 마이닝에 의해 발견되는 기존의 관념을 깨는 지식은 기업 경쟁력 강화에 결정적인 역할을 한다.[4)]

데이터 마이닝 기법 중 가장 활발하게 연구되고 있는 기법은 연관규칙 탐사기법이다. 연관규칙은 동시에 자주 나타나는 데이터들에 대한 연관성을 규칙의 형태로 표현한 것으로 이미 발생한 트랜잭션들에 대하여 데이터 사이의 연관성을 발견하여 이를 바탕으로 고객들의 구매 패턴을 분석하여 상품들의 시장성 예측 등에도 그대로 적용 할 수 있다. 연관규칙 탐사 분야는 응용성이 높아 기업의 마케팅, 판매전략, 고객 지원 등에 유용하게 이용되고 있다.[5)]

그러나 기존의 연관 규칙 탐사 기법은 데이터베이스에서 사용자가 지정한 정도의 통계적인 가치로 정의되는 수치를 만족하는 데이터들 사이에서 발생할 수 있는 연관성을 탐사하였다. 이러한 방법은 데이터베이스에 존재하는 각각의 데이터들이 모두 유사한 발생 빈도로 나타남을 전제로 하여 연관규칙을 탐사하는 방식이다. 그러나 실제로는 데이터베이스를 구성하는 데이터들은 데

---

3) Fayyad, U.M., G. Piatetsky-Shapiro, and P.Smyth, "From Data Mining to Knowledge Discovery," In Advances in Knowledge Discovery and Data Mining, Fayyad U.M, G.Piatetsky- Shapiro,P.Smyth and R.Uthurusamy, AAAI Press/Mit Press, CA., 1996, pp.1~34.

4) 이재규, 최형림, 김현수, 이경전, "전자상거래원론", 법영사, 1999.

5) 박종수, 유원경, 홍기형, "연관 규칙 탐사와 그 응용", 정보과학 학회지, 제16권,1998.

이터항목의 특성에 따라 상대적으로 빈번하게 나타나는 데이터도 존재하고 반대로 그렇지 못한 데이터도 존재한다. 그리고 빈번하게 나타나지 않는 데이터에도 경우에 따라서는 의미있고 중요한 정보가 존재할 수 있다. 기존 대부분의 연관규칙 탐사기법이 빈발하게 나타나는 데이터들만을 대상으로 탐사하는 알고리즘들이므로 희소하게 나타나는 데이터는 탐사할 수 없었다.[6] 희소한 데이터를 탐사 대상으로 하는 알고리즘들도 있으나 이 알고리즘들은 희소한 데이터의 중요도를 고려하지 않으므로 불필요한 정보를 탐사하는 문제가 있었다. 빈번하게 나타나지 않는 데이터를 탐사대상으로 하는 이유는 비록 희소하더라도 의미있고 중요한 정보를 찾아내려고 하는 것이므로 데이터의 중요성을 고려하면 더 효율적으로 탐사 할 수 있다.

본 논문은 희소한 데이터 중에서도 중요도를 고려하여 연관규칙을 탐사할 수 있는 알고리즘을 제안한다. 중요지지도를 고려하는 연관규칙 탐사기법은 불필요한 연관규칙들을 배제함으로써 기존의 방법보다 처리속도가 빠르면서 의미있고 중요한 연관규칙을 탐사할 수 있다.

6) R. Agrawal, Tomasz Imielinski, Arun Swami, "Mining Association Rules between Sets of Items in Large Database". Proccding of ACM SIGMOD, 1993.

2. 연구 내용 및 논문의 구성

본 논문에서는 데이터의 중요도를 고려하여 연관 규칙을 탐사하는 알고리즘을 제안한다. 희소한 데이터 중에서도 상대적으로 연관성이 있는 중요한 데이터들만을 대상으로 연관 규칙을 탐사하는 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다.

1장에서는 서론부분으로 연구배경 및 목적을 제시하여다.

2장에서는 데이터 마이닝의 기본 개념 및 응용 대해서 알아보았다.

3장에서는 연관규칙 탐사와 관련한 연구로서 기존의 알고리즘을 고찰하고, 문제점을 제시하였다.

4장에서는 문제점을 해결하는 방안으로 중요지지도를 고려하는 연관규칙 탐사 알고리즘을 제안하였다.

5장에서는 제안한 알고리즘을 시뮬레이션 방법을 이용하여 분석 평가 하였다.

6장에서는 결론과 향후 연구 방향을 기술하였다.

Ⅱ. 데이터마이닝 이론적 고찰

1. 데이터 마이닝 개념

대규모 데이터의 생성과 그것을 저장할 수 있는 기술 발달은 기업체나 연구기관 등에서 대규모 데이터베이스 구축을 가능하게 하였다. 그러나 최근에는 단순히 데이터를 확보하는 단계를 지나 확보된 데이터를 분석하여 실제 활동에 적용할 필요성이 증가 되었으며 이러한 요청에 의해 발생된 기술 분야가 KDD(Knowledge Discovery in Databses)[7]와 데이터 마이닝이다.[8]

데이터마이닝이란 자동화되고 지능을 갖춘 데이터베이스 분석기법으로 지식발견(KDD:Knowledge Discovery in Databses), 정보 발견(Information Discovery), 정보수확(Information havesting)등의 이름으로도 소개되어 왔다. 일반적으로 "대량의 데이터로부터 새롭고 의미있는 정보를 추출하여 의사결정에 활용하는 작업"이라 정의된다.[9] 용어에 '채굴하다'라는 의미로 'mining'을 포함시킨 이유는 데이터로부터 정보를 찾아내는 작업이 마치 금이나 다이아몬드를 발견하기 전에 수 많은 양의 흙과 잡석들을 파헤치고 제거하는 것과 유사하는 데에 기인하다.

1995년 캐나다 몬트리올에서 개최된 지식발견과 데이터마이닝에 관한 국제 학술대회(The first international conference on knowledge discovery & data mining)에서 지식발견은 데이터로부터 유용한 정보를 발견하는 프로세

---

7) Pieter Adriaans, Dolf Zantige, "Data Mining", Addison Wesley, 1996.
8) 이도현, "데이터 마이닝을 이용한 CRM" 정보과학회지, 제18권 11호, 2000, pp.4~11.
9) R. Agrawal and R. Srikant, "Fast algorithms for Mining Association Rules", Proc. VLDB, 1994.

스 전과정이라 정의하였다. 지식탐사 프로세스의 핵심적인 역할을 담당하는 데이터마이닝은 대량의 데이터로부터 패턴 인식, 통계적 기법, 인공지능 기법 등을 이용하여 숨겨져 있는 데이터간의 상호 관련성 및 유용한 정보를 추출하는 단계이다. 이 단계에서는 소비자 성향 파악을 통한 마케팅 및 판매 전략 수립, 고객 지원[10] 등과 같은 목적에서부터 의학 치료방법 및 범죄 방지에 이르기까지 여러가지 목적에 따라 알고리즘이 선택되고 원하는 결과를 얻기 위해 연속적으로 적용된다. 마이닝 결과 생성되는 유용한 지식은 의사결정에 직접적인 영향을 미칠 수 있다.

데이터마이닝은 데이터에 내재되어 있는 유용한 패턴이나 변수들간의 관계를 정교한 분석 모형을 사용하여 찾아내는 작업이다. 데이터 마이닝은 기업들이 보유한 기존의 경험적 지식을 인식하지 못했던 새로운 정보를 제공하여 경영의사결정에 도움을 준다.[11] 특히 데이터 마이닝에 의해 발견되는 기존의 관념을 깨는 지식은 기업경쟁력 강화에 결정적인 역할을 하다. 데이터 마이닝이 각 유통업, 은행업, 통신 산업, 보험업의 측면에서의 적용 사례는 아래와 같다.[12]

•유통업

유통업계에서의 데이터마이닝의 적용은 그들이 발행하는 신용카드와 전산화된 결제 시스템을 통해 축적된 고객들의 구매 정보에 대하여 이루어진다. 이러한 정보는 바구니 분석, 시계열 패턴 조사, 예측 모델의 개발 등에 적용된

10) 김재경, 이건창, 정남호, 권순재, 조윤호, "클레멘타인 데이터마이닝 솔루션을 이용한 웹 로그 분석", Information Systems Review, Vol. 4, No.1, 2002.3, pp.47~60.
11) 이도현, "데이터 마이닝을 이용한 CRM" 정보과학회지, 제18권 11호,2000, pp.4~11.
12) 이정원, 김호숙, 최지영, 김현희, 용환승, 이상호, 박승수, "데이터마이닝 알고리즘의 분류 및 분석", 정보과학회 논문지,데이터베이스 제 28권 제 3호, 2001.9, pp.279~299.

다. 바구니 분석은 고객들의 구매 행위 시 같이 구매되는 상품들을 발견하고 분석하여 상점의 진열 전략, 재고 전략, 판매촉진 등에 적용된다.

시계열 패턴(temporal sequence)조사는 시간에 따른 구매 행위에 대한 정보로 활용되며, 예측 모델은 고객의 구매 행위 등 특정 고객 집단을 목적으로 하는 효과적인 판매 전략을 세울 수 있다.[13)]

•은행업

은행에서 데이터마이닝은 사기 행위 색출(fraud detection), 고객 집단 분류(customer segmentation) 등의 분야에서 적용되고 있다. 사기 행위 색출은 데이터마이닝 기법을 이용하여 과거에 사기 행위로 판명된 신용카드 거래를 분석하여 사기 행위의 패턴 탐사를 통하여 이루어지며, 고객의 행위가 사기 행위 패턴과 유사할 경우 그 거래를 승인 하지 않는 시스템에 적용 되기도 한다.

고객 집단의 분류는 특정 고객 집단에만 특화된 차별화 서비스 이용에 적용된다.

•통신 산업

통신 회사들은 기존 고객을 유지하고 새로운 고객을 끌어들이기 위한 마케팅 전략으로써 데이터마이닝을 통한 통화기록 분석, 고객충성도(customer loyalty) 분석 등을 통하여 통화 사용 패턴을 이용한 가격 정책의 수립 등에 적용될 수 있다.

•보험업(Insurance)

---

13) 황현숙, 어윤양, "연관 마이닝과 고객 선호도 기반의 인터넷 상품 검색 시스템 설계 및 구현" 경영정보학 연구 제12권 제1호, 2002.3, pp.1~16.

보험 회사에서의 마이닝 기술 적용은 사기 색출, 상품 설계, 위험 분석 등에 적용된다. 분석된 결과는 보험 창구 패턴을 찾아내 보험 사기를 줄이는 작업이나 새로운 상품의 설계 등에 이용되며 보험 지급과 관련된 여러 요인의 분석을 통해 지급 부담 위험을 줄이는데도 사용된다.

2. 데이터 마이닝 기술

데이터마이닝은 데이터에서 더 많은 유용한 정보를 얻어내는데 도움을 주는 어떠한 방법이라도 유용한 것이다. 현재까지 알려진 수 많은 데이터 마이닝 알고리즘들은 각기 장단점을 가지고 있다. 그러므로 주어지는 데이터마이닝 작업의 목적에 따라 적합한 데이터마이닝 알고리즘을 선택하여 적용해야 한다.

데이터마이닝 방법은 목적에 따라 예측을 위한 것과 지식 탐사를 위한 것으로 구분한다. 예측은 특별한 목표에 관심을 가지고 과거의 기록을 기초로 미래의 새로운 사건에 적용하기 위한 방법이며, 분류, 시계열, 회귀분석 등의 방법이 활용된다.[14] 지식탐사는 예측을 포함하는 보다 포괄적인 개념이지만, 예측에 비하여 적은 경험정보로 작업이 가능하며 의사결정지원에 적합한 방법이다. 군집화, 연관규칙, 데이터베이스 분할, 요약화, 시각화, 편차탐지 등이 해당된다.[15]

이와 같은 목적별 방법들 중에서 사용자는 주어진 데이터마이닝 작업의 요구사항에 따라 적합한 선택이 필요한다. 최근 수년동안 학계, 연구계, 산업계

---

14) 정희택, "고객 관계 관리를 위한 워크플로우 시스템의 통합",정보과학회지, 제 18권 11호, 2000, pp.22~28.

15) 이정원, 김호숙, 최지영, 김현회, 용환승, 이상호, 박승수, "데이터마이닝 알고리즘의 분류 및 분석", 정보과학회 논문지, 데이터베이스 제 28권 제 3호, 2001.9, pp.279~299.

에서 데이터마이닝에 대한 연구가 이루져 왔다. 그간 제안된 다양한 데이터마이닝 기법들은 어떤 형태의 지식을 탐사하고자 하는가, 어떤 종류의 데이터베이스에 적용될 수 있는가, 어떤 분야의 기술에 바탕을 두고 있는가 등의 기준에 의거하여 아래와 같이 분류한다.[16]

1) 탐사될 지식의 형태에 따른 분류

•특성화(characterization)

데이터 집합의 일반적 특성을 분석하는 것으로 일반화 및 세분화 과정에 의한 자료 요약과정을 거쳐 특성 규칙을 발견한다.

•분류화(classification)

다른 클래스에 대한 차별적인 특성을 도출한다. 이와 같은 차별적인 특성은 소속 클래스를 알 수 없는 미지의 객체가 있을 때, 그 소속 클래스를 결정하는데 활용된다.

•군집화(clustering)[17]

유사한 특성을 갖는 데이터들을 묶음 지워주는 것이다. 인공지능 분야에서 분류는 감독학습임에 반해 클러스터링은 비감독학습으로 불린다. 감독학습이란 감독자가 자료를 집단별로 구분해 놓고 분류기준은 컴퓨터 프로그램이 학습에 의하여 발견하도록 하는 방법이다. 비감독학습은 감독이 없이 컴퓨터 프로그램 스스로가 자료 집단의 유사성을 바탕으로 집단을 분류하는 방식이다.

---

16) 김정자, 이도현, "데이터 마이닝 기술 및 연구동향", 정보과학학회지, 제 16권 제 9호, 1998.9, pp.6~14.

17) 황인수, "고객관계관리에서 신경망을 이용한 제품-고객군의 형성에 관한 연구". 경영정보학연구, 제 11권 제 4호 , 2001, pp27~41.

•연관규칙 탐사(association rule mining)[18]

여러 개 트랜잭션들 중에서 동시 발생하는 트랜잭션의 연관관계를 발견하는 것이다. 경향 분석, 패턴 분석에 사용한다.

•경향 분석(trend analysis)

시계열 데이터(주식, 물가, 판매량, 과학적 실험 데이터)들이 시간축으로 변하는 전개과정을 특성화하여 동적으로 변화하는 데이터의 분석을 수행한다.

•패턴 분석(pattern analysis)

대용량 데이터 베이스 내의 명시된 패턴을 찾는 것이다.

2) 탐사될 데이터베이스의 타입에 따른 분류

탐사될 데이터 베이스가 관계형 데이터베이스인지 혹은 객체지향형, 네트워크형, 계층형인지에 따라서 적용할 수 있는 데이터마이닝 기법이 달라지게 된다. 최근 POS 시스템에서 흔히 사용되는 트랜잭션 데이터베이스도 이와 같은 구분에 따라 고려 할 수 있다. 한편 공간 데이터베이스, 멀티미디어 데이터베이스 처럼 문자나 숫자가 아닌 복잡한 자료만을 포함하고 있는 경우에도 새로운 데이터마이닝 기법이 필요하게 된다.[19]

---

18) Chang, G, Healey, M. J., McHugh, J. A. M., and wang, J. T. L., "Mining the World Wid Web: An Information Search Approach", Kluwer Academic publishers, 2001.

19) 남도원, 김성민, 이동하, 오재훈, 김성훈, 이전영, "한시적 연관규칙에서의 부분 구간 탐사", 한국정보과학회 데이터베이스 학술대회, 1999.

3) 적용기술의 종류에 따른 분류

데이터로 사용된 변수의 분포, 상관관계, 탐사된 규칙에 대해 확신하기 위한 수단으로 통계를 이용하고, 논리를 이용한 기호학습 방법도 사용한다. 신경망은 분류화를 위한 기법으로 많이 사용하며, 수행 결과를 보다 효과적으로 보여주기 위해 가시화 기법을 사용하기도 한다. 일반적으로 이러한 기법들은 독립적으로 사용되기 보다는 혼합적으로 사용되는 경우가 많다.[20]

3. 연관규칙

1) 빈발항목집합의 정의

트랜잭션이 빈번하게 발생하는 소매점의 물품 판매에서 만들어지는 트랜잭션 데이터베이스를 고려해보자. 항목들(예를 들면, 소매점에서 판매된 물품 항목들)의 집합 $I=\{i_1, i_2, i_3..., i_m\}$이 주어지면, 트랜잭션 T는 I의 부분집합으로 정의된다($T \subseteq I$). 트랜잭션이라 함은 구매자가 소매점에서 한꺼번에 구매하는 물품들의 집합으로 볼 수 있다. 집합과 같이 트랜잭션들은 중복된 항목을 허용하지 않는다. 그러나 우리는 순수한 집합의 개념을 확장하고 트랜잭션과 다른 모든 항목집합들 내에 있는 항목들은 정렬된 것으로 가정한다. 데이터베이스 D를 n개의 트랜잭션들의 집합이라 하고 각 트랜잭션은 고유한 트랜잭션 번호 (TID)가 부여 된다. 만일 트랜잭션 T가 X의 모든 항목들을 포함한다면 ($X \subseteq T$), T가 집합 X(물론, $X \subseteq I$)를 지지한다(support)고 한다. 우리는 X의 지

20) Jiawei Han and Micheline Kamber, "Data Minig: Concepts and Techniques", Morgan Kaufmann Pubishers, 2001.

지도를 생략된 형태 supp(X)로 정의하며 이는 X를 지지하는 D에 있는 모든 트랜잭션들의 개수를 의미한다. 만일 사용자가 정한 최소지지도 S_{min}에 대하여 $supp(X) \geq S_{min}$이라면, 집합 X는 빈발하다(이런 경우에 항목들의 집합 X를 일반적으로 large itemset이라 하고 또는 frequent itemset이라고도 한다)고 한다. 최소지지도를 사용하는 이유는 D에 대하여 관심있을 정도로 빈발하게 나타나는 항목만을 고려하기 위함이다. 항목 집합 X의 개수를 k=|X|로 나타내고 이를 k-항목집합이라 부른다.[21]

다음은 빈발 항목집합을 효과적으로 찾는 과정에서 얻어진 특성들로 대부분의 연관규칙 탐사 알고리즘에서 사용되고 있다.

- 특성1(부분집합의 지지도): 만일 항목 집합 A, B에 대하여 $A \subseteq B$이면, B를 지지하는 D의 모든 트랜잭션들이 필연적으로 A 또한 지지하므로 $supp(A) \geq supp(B)$이다.
- 특성2(빈발하지 않은 집합들의 상위집합들은(supersets)은 빈발하지 않다): 만일 항목집합 A가 D에서 최소지지도에 미치지 못한다면, 즉 $supp(A) < S_{min}$, 특성 1에 의하여 $supp(B) \geq supp(A) < S_{min}$이기 때문에 A의 모든 상위 집합 B는 빈발하지 않을 것이다.
- 특성3(빈발 항목집합들의 부분집합들은 빈발하다): 항목집합 B가 D에서 빈발하다면, 즉 $supp(B) \geq S_{min}$, 특성 1에 의하여 $supp(A) \geq supp(B) \geq S_{min}$이므로 B의 모든 부분집합 A는 D에서 또한 빈발한 것이다. 특히, 만약 $A=\{i_1, i_2, i_3, \cdots, i_k\}$가 빈발하면, 그것의 모든 k개의 (k-1)-부분집합들도 빈발하다. 그 역은 성립하지 않는다.

21) 최영희, 장수민, 유재수, 오재철, "수량적 연관규칙탐사를 위한 효율적인 고빈도 항목열 생성기법", 한국정보처리학회논문지, 제6권 제10호, 1999, pp.2597-2607.

2) 연관규칙의 정의

X와 Y를 항목들의 집합이라 하자. 연관규칙(association rule)은 R: X → Y 형식의 함수이고, 이때 X와 Y는 서로 같은 원소를 갖지 않는 항목집합이다: X, Y⊆I이고 X∩Y=∅ 이고, Y≠∅ 여야 한다. X를 규칙의 조건부(antecedent)라 하고 Y를 결과부(consequent)라 한다. 만일 한 트랜잭션이 X를 지지한다면, 또한 어떤 확률에 의해 Y도 지지할 것이라는 예측으로 이해될 수 있는 것이 연관 규칙이다. 이런 확률을 이 규칙의 신뢰도 (conf(R)로 표시)라 한다. R의 신뢰도는 X를 지지하는 T에 대하여 Y 또한 지지할 조건부 확률로 정의된다. 즉, conf(R)=supp(X∪Y)/sup(X), D에 있는 규칙 R에 대한 지지도는 supp(X∪Y)로 정의 한다. 규칙의 신뢰도는 얼마나 조건부에 대하여 결과부가 자주 적용할 수 있는지를 나타내고 반면 지지도는 그 규칙 전부가 얼마나 믿을 만한지를 보여준다. 규칙이 데이터베이스에서 적절해지려면 충분한 지지도와 신뢰도를 가져야 한다. 그러므로 어떤 주어진 최소신뢰도 C_{min}과 최소 지지도 S_{min}에 대하여 만일 $conf(R) \geq C_{min}$ 이고 $supp(R) \geq S_{min}$하면 규칙 R은 D에 대하여 성립한다.

Ⅲ. 연관규칙 탐사에 관한 선행 연구

1. 연관규칙 탐사 기법

연관규칙 탐사문제는 Agrawal[22]에 의해 처음 소개된 이후로 다양한 문제 영역이 제시되었고, 새로운 알고리즘이 지속적으로 소개되고 있다. 그러나 다른 알고리즘에도 불구하고 이들의 기본적인 스키마는 유사하다. 즉, 연관규칙을 탐사하기 위하여 데이터 베이스에 있는 모든 빈발 항목들의 지지도를 계산하여 빈발 항목집합을 찾고, 이로부터 주어진 신뢰도를 바탕으로 실제의 규칙을 탐사하는 2단계 과정으로 이루어진다.[23]

첫번째 단계는 빈발 항목집합을 찾는 단계로 주어진 최소지지도(S_{min})이상의 트랜잭션 지지도를 가지는 항목집합들인 빈발 항목집합을 찾는 과정이다.

두번째 단계 첫번째 단계에서 생성된 빈발 항목집합을 사용하여 연관규칙을 생성하는 단계이며, 최소신뢰도 이상의 규칙을 찾는다. 빈발 항목집합 L에 대하여 L의 모든 공집합이 아닌 부분집합을 찾는다.

연관규칙 탐사의 전체성능은 첫 번째 단계에서 결정된다. 데이터베이스 속에 고려대상 빈발 항목의 수는 모든 항목들의 멱집합(power set)의 크기와 같다. 즉 항목들의 수 증가에 대하여 고려해야 할 항목의 크기는 기하급수적으로 증가하여 상당량의 처리시간과 메모리를 요구한다. 따라서 이러한 문제의 해

---

22) Fayyad, U.M., G. Piatetsky-Shapiro, and P.Smyth, "From Data Mining to Knowledge Discovery," In Advances in Knowledge Discovery and Data Mining, Fayyad U.M, G.Piatetsky- Shapiro,P.Smyth and R.Uthurusamy, AAAI Press/Mit Press, CA., 1996, pp.1~34.

23) 박정호, "Apriori 알고리즘 연관 규칙마이닝 기법을 이용한 정보검색", 고려대학교 대학원 석사 논문, 1999.

결을 위한 연구로는 Apriori를 근거로 성능을 개선시킨 AprioriTid, AprioriHybrid, DHP등이 있었으며, 또한 Apriori와는 다르게 접근한 Partition, DIC, Direct Sampling, Sampling Approach 등의 연구도 있었다.[24] 연관규칙 탐사의 대표적인 알고리즘으로 알려진 Apriori는 해당 항목의 발생 유무만 고려하는 이진항목의 탐사 알고리즘이다. 그리고 이 알고리즘의 탐사 원리가 간편하고 이해가 용이하여 많은 알고리즘에서 응용하고 있다.

Apriori 알고리즘 외에도 범주적 속성(categoritical attribute) 데이터 뿐만 아니라 수량적 속성(quantitative attribute) 데이터의 연관규칙을 찾는 알고리즘, 계층도(taxonomy:is-a hierarachy)를 이용하여 일반화된 연관규칙을 찾는 알고리즘,[25] 항목간의 순차적 특성(sequential attribute)을 찾는 알고리즘,[26] 갱신과 유지(update and maintenance)를 위한 알고리즘, 마케팅의 세일즈 데이터 뿐만 아니라 센서스 데이터와 같이 상대적으로 패턴의 길이가 긴 데이터에 대하여 연관규칙을 찾는 알고리즘[27] 그리고 항목간의 주기적 특성을 (cyclic attributes)을 찾는 알고리즘 등이 있다.[28]

이러한 수 많은 알고리즘 중에서 상향식 알고리즘으로 Apriori, 하향식으로 Max-Miner를 들수 있으며, Apriori를 보완하는 방법으로 MSApriori 알고리즘과 의미 있는 희소 데이터를 포함한 연관 규칙 탐사 기법인 MSApriori, RSAA 등이 제안 되었다.

---

24) 이정원, 김호숙, 최지영, 김현희, 용환승, 이상호, 박승수, "데이터마이닝 알고리즘의 분류 및 분석", 정보과학회 논문지,데이터베이스 제 28권 제 3호, 2001.9, pp.279~299.

25) 황정희, 신예호, 류근호, "트리거와 점진적 갱신기법을 이용한 연관규칙탐사의 능동적 후보 항목 관리 모델", 정보과학회논문지, 데이터베이스 제 29권 제1호, 2002.2, pp.1~13.

26) R. Agrawal and R. Srikant, "Mining Sequential Patterns", Proc. ICDE, March 1995.

27) Witten, I.H., and Frank, E., "DataMining: Practical Machine LearningTools and Techniques with Java Implementations", Morgan Kaufmann Pubishers, 2000.

28) 안효성, "연관 규칙을 활용한 데이터베이스 지식 탐색 도구 구현에 관한 연구", 국민대학교 대학원 석사 논문, 1999.

1) Apriori 알고리즘

Apriori 알고리즘[29]은 데이터베이스에서 후보항목집합을 구성하고, 구성된 후보항목집합에서 빈발항목집합을 탐사하는 과정으로 구성된다. Apriori 알고리즘은 후보항목 생성 시 모든 데이터베이스에서의 데이터 항목에 대한 생성이 아닌, 전 단계의 빈발항목집합을 대상으로 후보항목을 구성한다. 빈발 항목집합(Large Item Set)은 사용자가 정한 최소 지지도 minsup에 대하여 데이터 항목 집합 X의 지지도 sup(X)가 minsup 보다 크면 X를 빈발하다(large)라고 정의한다. 빈발 항목은 빈발 항목집합의 원소에 포함되는 항목을 의미한다. K-빈발항목집합은 k개의 데이터항목으로 구성된 빈발항목집합으로 L_k로 표현한다. 후보항목 집합(Candidate set)은 빈발항목집합의 원소가 될 가능성이 있는 항목들로 구성된 집합으로 빈발항목 집합을 탐사하기 위해 사용되는 집합이다. K-후보항목집합은 k개의 데이터항목으로 구성된 후보항목을 말하며 C_k로 표현한다. k-항목 집합(k-itemset)은 항목 집합의 원소가 k개의 데이터로 구성된 항목 집합이다.

Apriori 알고리즘은 전 단계에서의 빈발항목집합에서 현재 단계의 후보항목집합을 구성하고 난 후, 데이터베이스의 스캔을 통해 후보항목집합의 지지도를 계산하고, 사용자가 정의한 최소지지도를 기초로 하여 현 단계의 빈발항목집합을 구성한다. Apriori 알고리즘의 단계의 진행은 데이터 항목의 증가에 따라 반복적으로 진행된다. K 단계에서의 Apriori의 빈발 항목 탐사는 k-후보항목집합에 대하여 각각의 지지도를 셈한 후 이들 중에서 지지도를 만족하는 항목의 탐사를 통해 이루어진다. Apriori는 더 이상의 후보항목을 생성 할

29) R. Agrawal and R. Srikant, "Fast algorithms for Mining Association Rules", Proc. VLDB, 1994.

수 없을 때까지 반복되어 빈발항목을 탐사한다.

후보 항목집합의 구성은 전 단계의 빈발 항목집합의 조인(join)연산과 전지 과정을 통해 구성된다. 조인연산은 두 집합의 곱집합을 구한 것과 같다. 전지 과정은 조인을 통해 생성된 후보항목집합의 부분집합이 전 단계의 빈발 항목집합의 원소가 아닌 경우, 그 항목을 삭제하는 과정이다. 그 이유는 전 단계에서 빈발하지 못한 항목은 다음 단계에서도 빈발하지 못하기 때문이다. 전지 과정은 불필요한 후보항목의 수를 줄여 데이터베이스를 읽는 회수를 감소 시키기 위하여 추가된 과정이다.

```
//DB검색하여 C1, L1 생성한다.
L1 = {large 1-itemsets}
For (k=2; Kk-1 ≠ Ø ;k++) do begin
    Ck =Apriori-Gen(Lk-1);// New candidates
   for all transation t ∈ D do begin
   Ct = subset(Ck, t);
     for all candidates c ∈ Ct do
           c.count++;
  end
    Lk = {c ∈ Ck| c.count ⊇ Smin}
end
Answer = Uk Lk;
```

<알고리즘 1>Apriori 알고리즘

1)결합(join)단계

insert into C_k

select p.item$_1$, p.item$_2$, p.item$_3$,… p.item$_{k-1}$, q.item$_{k-1}$

from L_{k-1} p, L_{k-1} q //self join

where p.item$_1$= q.item$_1$, q.item$_2$, … p.item$_{k-2}$

= q.item$_{k-2}$,p.item$_{k-1}$ ⊂ q.item$_{k-1}$;

2)전지(prune) 단계

forall itemset c ∈ C_k do

forall (k-1) - subsets s of c do

if (s ∈ L_{k-1}) then delete c from C_k

<알고리즘 2> Apriori-gen 함수

<알고리즘 1>은 Apriori 알고리즘이며, <알고리즘2>는 알고리즘 내에서 항목의 결합(join)과 전지(prune)를 수행하는 Apriori-Gen 함수이다. <알고리즘 1>에서 C_k는 후보 k-항목집합(k-candidate itemset)이고, L_k는 빈발 k-항목집합을 나타낸다. 알고리즘의 첫 번째 단계에서 빈도수를 계산하여 빈발 1-항목집합 L_1을 결정하고, k(k≥2) 번째는 두 단계로 분할하여 알고리즘이 진행된다.

먼저, (k-1) 번째 검색에서는 발견된 빈발 항목집합 L_{k-1}를 이용하여 후보 항목집합 C_k로 만든다. 다음으로 DB를 검색하여 C_k에 있는 후보 항목 집합의 지지도를 계산한다. C_k에 있는 후보 항목집합 중에서 최소지지도를 만족하는 항목만 선택하여 L_k에 진입시킨다. 이러한 시행은 L_k가 더 이상 발견되지 않을 때까지 반복한다. 이 <알고리즘 1>의 성능은 <알고리즘 2>의 결합(join)과 전지(prune)단계에 많은 영향을 받고 있다. Apriori 알고리즘의 탐사

과정의 예를 들어보자. 주어진 데이터베이스 D는 <표3-1>과 같이 4개의 트랜잭션과 5개의 항목으로 구성되었다.

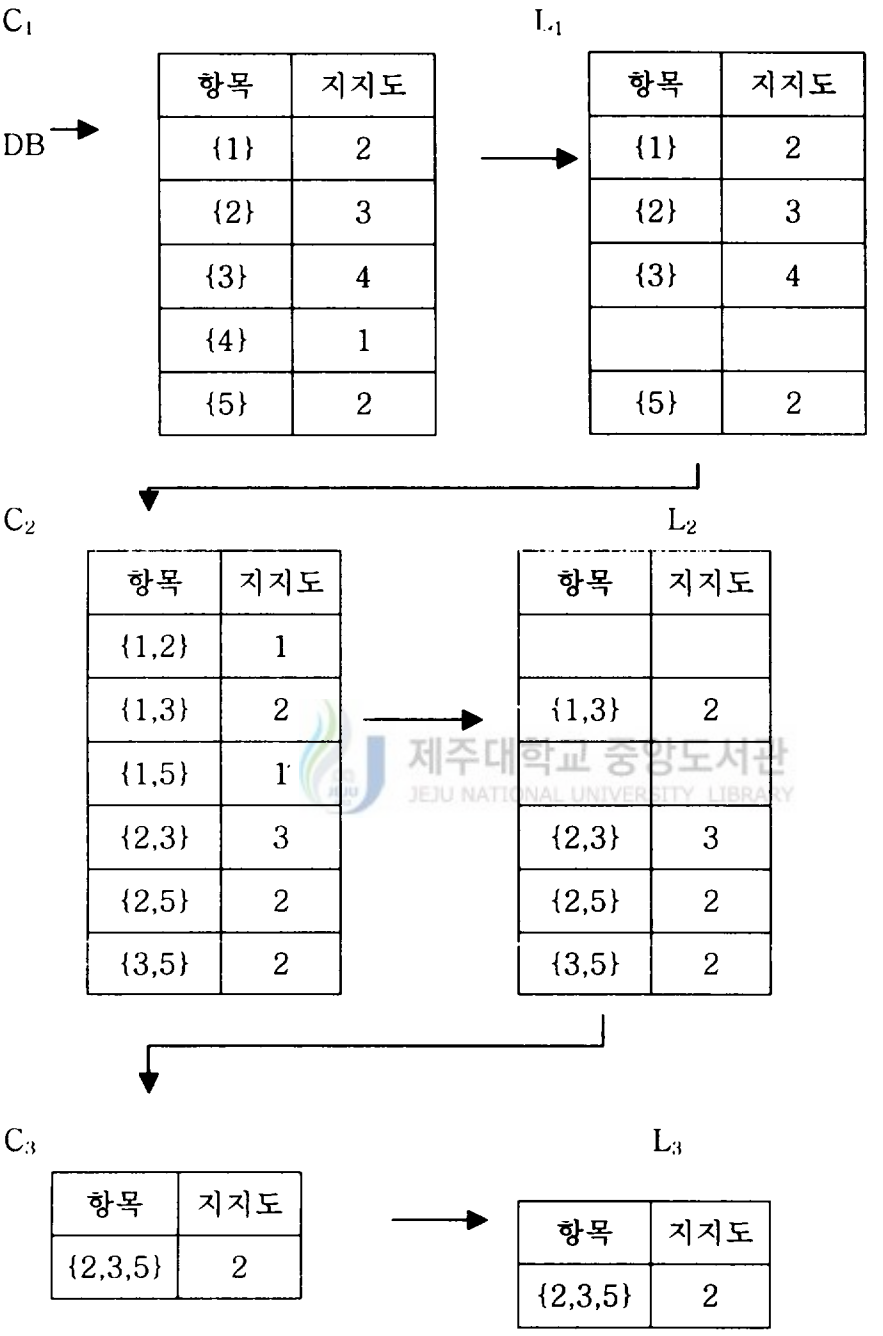
<표 3-1> Apriori 데이터 베이스 예

TID	항목 (items)
1	1,3,4
2	2,3,5
3	1,2,3,5
4	2,3

<알고리즘2>의 결합단계에 의해서 L_3 = {{1,2,3},{1,2,4},{1,3,4},{1,3,5}, {2,3,4}} 일때, 후보 4-항목집합 L_4은 L_3의 자기 결합(self join)에 의해서 {{1,2,3,4},{1,3,4,5}}이 생성 된다.

전지단계에서는 결합단계에서 생성된 결과의 각각에 대하여 부분집합이 L_3에 존재유무를 평가하여 존재하지 않는 집합은 전지하는 단계이다. {1,2,3,4}의 3-부분집합 = {{1,2,3},{1,2,4},{1,3,4},{2,3,4}}인데, 3-부분집합 중에서 {1,4,5},{3,4,5}는 L_3의 원소가 아니므로 {1,3,4,5}는 전지한다.

<그림 3-1>은 <표3-1> 예제 데이터베이스에 대하여 Apriori 알고리즘으로 빈발항목집합을 탐사하는 과정이다. <표 3-1>에는 트랜잭션 4개, 항목 5개의 예제 데이터베이스이며, 최소지지도는 50%로 가정한다. 그러므로 트랜잭션 4개중에서 2개 이상에 해당 항목이 포함되어 있어야 빈발하다고 할 수 있다.



<그림 3-1> Apriori 수행과정

<그림 3-1>을 보면, 첫 번째 데이터베이스 검색에서 각 항목의 지지도를 계

산하고, 그 중에 최소지지도를 만족하는 항목만을 L_1에 진입시킨다. <그림 3-1>의 후보 1-항목집합 C_1에는 데이터베이스를 검색하여 계산된 항목 {1},{2},{3},{4},{5}의 지지도가 포함되어 있다. 각 항목의 지지도는 2, 3, 4, 1, 2 이다. 여기서 항목 {4}의 지지도는 1이므로 최소지지도 50%(2)를 만족하지 못한다. 따라서 항목 {4}는 제거된 후, 잔여 항목들을 빈발 1-항목집합 L_1에 진입시킨다. L_1의 항목으로 결합단계와 전지단계를 거쳐 후보 2-항목집합 C_2를 만든다. C_2는 L_1항목인 {1,2},{1,3},{1,5},{2,3},{3,5}를 생성한다. 그리고 다시 데이터베이스를 검색하여 C_2 의 지지도를 계산하여 최소지지도를 만족하는 후보 항목집합만을 L_2에 진입시킨다. 이런 과정을 반복하여 L_3를 만들고, L_4가 공집합이므로 알고리즘의 진행을 종료한다. 최종적으로 사용자가 얻을 수 있는 빈발 항목집합은 다음과 같다.

L_1 = {{1},{2},{3},{5}}, L_2 = {{1,3},{2,3},{2,5},{3,5}},

L_3 = {{2,3,5}}

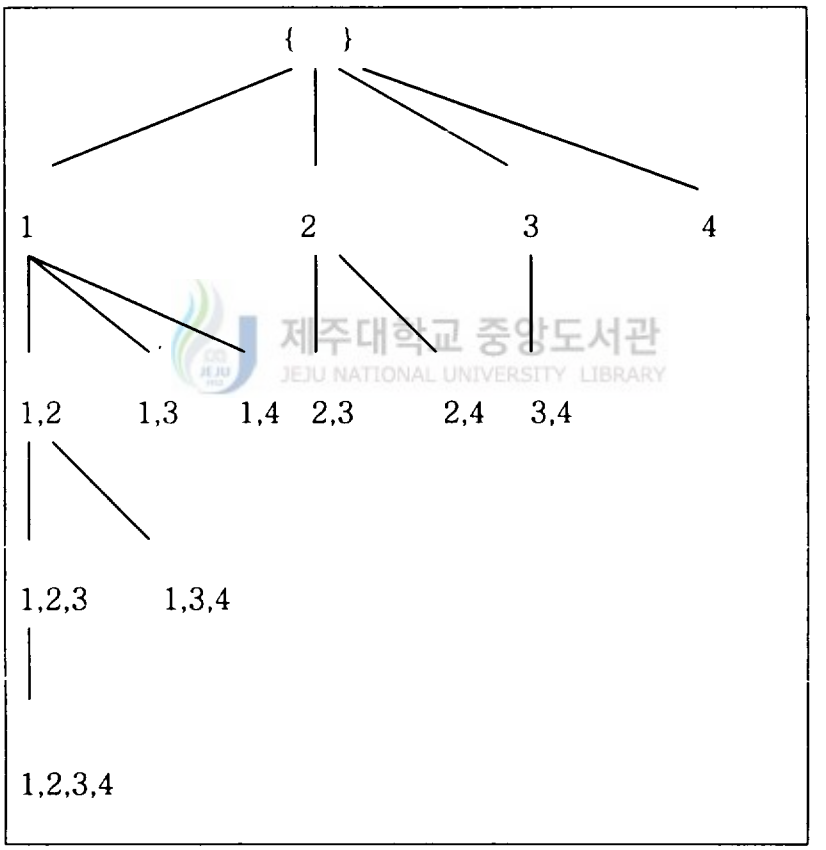
예를 통하여 알 수 있듯이 Apriori알고리즘은 항목의 존재 유무와 발생 빈도만으로 빈발 항목집합을 생성하며 뿐만 아니라 해당 항목의 중요도 또한 고려하지 않는다. 또한 Apriori는 1-항목집합에서 k-항목집합으로 항목을 점차 확장해 가는 상향식으로 빈발 항목집합을 생성하는 알고리즘이다. 이와 반대인 하향식 알고리즘은 Max-Miner[30] 를 들 수 있다.

2) Max-Miner 알고리즘

Max-Miner는 Apriori와는 달리 예상되는 최대 길이의 항목 집합의 빈발여부를 확인하고 길이를 감소시키는 하향식(top-down) 방법이다. 처음에는 최

30) 박원환, "수량 연관규칙을 위한 데이터 마이닝 알고리즘에 관한 연구",순천향대학교 대학원 석사 논문, 2001.

대 길이의 항목집합을 고려한다. 이 항목이 빈발하지 않을 경우에만 부분 항목집합을 만들어 다시 DB를 검색하여 빈발여부를 확인하다. Max-Miner는 집합열거 트리(set-enumeration tree)를 사용하여 항목들을 표현함으로써 보다 효과적인 전지전략(pruning strategy)을 수행토록 하고 있다. <그림 3-2>는 집합열거 트리 예이며, 각 항목의 순서에 따라 부모 노드(parent node)와 자식노드(child node) 관계이며, 사전적 순서를 가진다.



<그림3-2> 항목 4개 집합열거 트리

위 집합열거 트리는 항목집합을 어떻게 완전하게 열거하는지를 나타내며, 탐색은 너비-우선(breadth-first)방법을 사용한다. 그러므로 만약 전지가 없다면, 집합열거 트리는 빈발 항목집합을 탐사하기 위하여 모든 항목집합을 고

려해야한다. Apriori는 빈발하지 않는 항목집합의 부분 항목집합만 전지하지만, Max-Miner는 빈발하지 않는 항목 집합의 부분집합 뿐만 아니라 빈발 항목집합의 상위집합 (superset)의 전지에도 사용된다. 집합열거함수의 각 노드는 후보그룹(candidate group)을 표현한다. 후보그룹 g는 두개의 항목집합, 즉 머리(head)와 꼬리(tail)로 구성되며, 각각은 h(g), t(g)로 표현한다. 머리는 집합열거 트리의 노드에 의해 항목집합을 열거하며, 꼬리는 순서적 집합(ordered set)이며 이후 단계의 하위노드(sub-node)에는 나타나지만, h(g)에는 포함되지 않는 모든 항목들을 포함한다. 예를 들면, <그림 3-2>에서 항목집합 {1}이고, $t(g_1)$= {2, 3, 4}이며, 항목집합 {2}의 $h(g_2)$ = {2} 이고, $t(g_2)$ = {3, 4} 등이다. 후보그룹 g에 대한 지지도는 h(g), h(g)∪t(g)와 h(g)∪{i}(모든 i ∈ t(g))를 계산하여 최소지지도를 만족하지 못하면 제거한다.

```
Max-Miner(Data-sat T)

//Returns the set of maximal frequent itemsets present in T

    Set of Candidates Groups C ← {}
    Set of itemsets F ← (Gen-Initial-Groups(T,C))

  While C is non-empty do
    scan T to count the support of all candidate group in C
      for each g ∈ C such that h(g)∪t(g) is frequent do
          F ← F∪{h(g)∪t(g)}
      Set of Candidate Groups C_new ← {}
      for each g ∈ C such that h(g)∪ t(g) is frequent do
```

```
        F ← F ∪ {Gen-Sub-Nodes(g, C_new)}
    C ← C_new

remove from F any itemset with a proper superset in F
remove from C any group g such that h(g)∪t(g)
        has a superset in F
Return F
```

<알고리즘 3> Max-Miner알고리즘

<알고리즘 3> Max-Miner알고리즘의 최소지지도도 Apriori 와 같이 사용자에 의해 정의된다. While 루프는 집합열거 트리를 너비-우선 탐색 방식으로 탐색하도록 구현된다. 초기 후보그룹 생성(Gen-Initial-Groups) 함수는 DB의 모든 트랜잭션을 검색하여 항목 도메인과 트리에서 2단계 레벨의 탐색 기반을 마련하는 작업을 수행한다. 첫째 시행에서 빈발 1-항목집합 F_1에 있는 각 항목 i에 대하여 h(g)에 i를 입력하고, i 이후의 항목들은 t(g)에 입력하여 후보그룹 g를 만들고, g를 후보 그룹집합 C에 진입시킨다. Fi을 최대길이의 항목집합만을 저장하여 Max-Miner로 리턴한다. Max-Miner에서 항목의 순서화 (재순서화)작업은 상위 빈발 항목 전지작업의 효율성을 증가 시키기 위하여 이루어진다. 후보그룹 g에 대하여 h(g)∪t(g)가 빈발할 때 상위빈발함수 전지 작업이 이루어진다. 순서화된 항목들 중에서 제일 마지막에 나타나는 항목이 대부분의 후보그룹에 나타난다.

예를 들면, <그림 3-2>의 항목 4는 각 노드의 머리나 꼬리에 나타나고 있다. 그러므로 대부분의 빈발 항목 위치는 항목순서의 끝에 나타난다. 이러한 순서화 작업은 두개의 함수, 즉 초기 후보그룹 생성함수와 서브노드 생성 함수에서 시행된다.

Max-Miner는 초기 후보그룹 생성함수에서 리턴한 F의 값으로 구성된 집합 열거 트리의 각 후보그룹 g의 지지도를 계산한다. 지지도는 h(g) ∪ t(g)와 h(g) ∪ {i} (i ∈ t(g)) 두가지 경우에 대하여 계산하며, 각각이 같은 트랜잭션에 존재할 때 지지도를 증가시킨다. 이렇게 지지도 계산이 끝나면, 지지도에 따라 빈발여부를 판단하여 다음 사항을 수행한다.

a) 만약 g의 h(g)∪ t(g)가 빈발하며, h(g)∪ t(g)를 F에 저장한다.

b) 만약 g의 h(g)∪ t(g)가 빈발이 아니며, h(g) u {i } (i∈ t(g))중에서 빈발이 아닌 항목을 t(g)에서 삭제한 후에 잔여 항목들만으로 순서를 재결정한다. t(g)에서 가장 높은 지지도 항목 i 와 h(g)의 합집합을 F에 진입시켜 새로운 그룹집합 C_{new}를 만든다. 새로운 항목그룹 g'를 만들기 위해 각 항목 i(i∈t(g))를 h(g)에 추가하여 g'의 h(g')입력한다g'의 t(g')에는 t(g')에는 t(g)의 i 이후의 항목들을 입력하여 새로운 후보그룹 g'를 만들어 C_{new}에 진입시킨다. C_{new}에 있는 g 중에서 h(g)∪t(g)가 항목집합 F을 부분집합이면 C_{new}에서 삭제한다. B) 의 경우는 서브노드 생성 함수에서 수행한다.

Max-Miner로 리턴된 값인 C_{new}는 C에 저장하여 동일함 방법으로 C가 더 이상 발견되지 않을 때까지 반복한다. 이들의 특징은 빈발항목집합의 탐사시간은 후보 항목집합으로 구성된 탐사공간의 크기와 지지도 계산을 위한 데이터베이스 검색에 따라 결정됨을 감안할 때, Apriori는 후보 항목집합 생성은 함수 Apriori-gen의 결합과 전지 단계에서 이루어지며, 진행방법도 1-항목에서 L항목으로 점차적으로 확대하는 상향식이다. 항목길이 1의 확장은 한번의 데이터 베이스 검색을 요구하므로 항목수가 적은 응용에서는 효율적이지만, 항목수가 많을 경우는 탐색공간의 크기가 기하급수적으로 증가하는 문제점이 있다. 즉 항목열의 수가 L이면, 2^{l} 개의 부분 집합이 생성해야 하고, 또한

항목 집합이 빈발 여부 확인을 위해 데이터베이스의 검색이 필요하다.

반면에 Max-Miner는 예상되는 최대 길이 항목 집합의 빈발여부를 확인하여 길이를 감소시키는 하향식 방법을 사용하고 있다. 처음에 예상되는 최대길이의 항목집합에서 출발하여 빈발일 경우에는 그 항목집합의 부분 항목집합을 더 이상 고려하지 않지만, 빈발이 아닐 경우에는 그 항목의 부분 항목집합을 만들어 다시 데이터 베이스를 검색한다. 따라서 항목수가 많은 경우에는 Max-Miner알고리즘이 보다 효과적이 될 수 있다.

3) MSApriori (Multiple Support Apriori)알고리즘

Apriori 알고리즘은 최소지지도와 최소신뢰도를 만족하는 모든 데이터들을 찾는 방법으로 전체 데이터베이스에 대해 단일한 최소지지도가 사용된다. 즉 Apriori 알고리즘은 데이터베이스에 존재하는 모든 데이터들은 유사한 빈도수를 가지고 있는 것으로 가정하고 연관규칙을 탐사하는 방법이다. 그러나 실세계의 많은 응용에서 모든 데이터들이 유사한 발생 빈도수를 가지고 나타나는 경우는 드물며 상대적으로 많은 빈도를 가지고 나타나는 데이터들이 존재하는가 하면 그렇지 않고 상대적으로 희소한 빈도를 가지고 나타나는 데이터들도 있다.

Apriori 알고리즘에서 최소지지도를 큰 값으로 설정한 경우, 희소한 데이터들에 대한 규칙은 탐사할 수 없다. Apriori 알고리즘에서 상대적으로 희소한 빈도를 갖는 데이터들에 대한 연관규칙을 탐사해야 하는 경우, 사용자는 최소지지도를 낮게 설정하여야 한다. 그러나 최소지지도를 작은 값으로 설정한다면 희소한 데이터 뿐만 아니라 그 최소지지도를 만족하는 빈발하는 데이터들로 구성된 모든 규칙들이 부가적으로 탐사되어 작은 최소 지지도를 만족하는 모든 데이터들이 서로 연관성을 가지고 있는 것으로 탐사하는 문제가 발생한다. 따라서 데이터들의 빈도 형태를 획일적으로 가정하는 Apriori 알고리즘을

개선한 방법인 MSApriori(Multiple Support Apriori)[31] 알고리즘이 제안되었다.

MSApriori알고리즘은 데이터들의 빈도 형태를 고려하기 위하여 데이터 각각의 최소지지도인 MIS(minimum item support)를 사용한다. 데이터베이스를 구성하는 각각의 데이터들은 자신의 MIS 값을 가지고 있다. MIS는 각각의 데이터항목에 지정된 최소지지도이며 데이터항목 i 의 MIS는 MIS(i) 로 표현한다. 데이터의 MIS는 다음과 같은 방법에 의해서 계산된다.

$$MIS(i) = M(i) \quad \text{if } (M(i) > Ls)$$
$$Ls \quad \text{Otherwise}$$
$$M(i) = ß \times f(i) \quad (0 \leq ß \leq 1)$$

MSApriori 에서도 최소한의 지지도인 LS를 만족하는 데이터를 탐사대상으로 하며 ß라는 값을 각각의 데이터의 빈도수 f(i)에 곱한 후 그 값을 LS와 비교하여 데이터들의 MIS를 구하게 된다.

MSApriori 알고리즘에서의 최소 지지도는 규칙을 구성하는 데이터 항목 중에서 가장 낮은 MIS값이 그 규칙의 최소 지지도로 이용된다. 즉 규칙 R : $i_1, i_2, i_3, \cdots, i_k, i_{k+1}, i_r$ 인 규칙에 대한 최소 지지도는 $\min(MIS(i_1), MIS(i_2), \cdots, MIS(i_r))$으로 설정된다. MSApriori 알고리즘은 MIS를 이용하기 때문에, 규칙을 구성하는 항목들이 빈발항목들로만 구성되었다면 그 규칙은 비교적 높은 최소 지지도에 의해 탐사되며, 반대로 희소한 데이터들로만 구성된 규칙의 경우에는 비교적 낮은 최소지지도가 적용되어 규칙을 탐사한다. MSApriori는 이러한 방법으로 Apriori 알고리즘에서의 일괄적인 지지도 적

31) Bing Liu, Wynne Hsu, Yiming Ma, "Mining Association Rules with Multiple Minimum Supports", Proceedings of the ACM SIGKDD(KDD-99), 1999.

용으로 생기는 문제점을 해결한다.

예를 들면 데이터베이스에 bread, shoes, clothes라는 데이터들이 존재하고 사용자가 지정한 이들의 MIS는 아래와 같다면,

MIS(bread)= 2% MIS(shoes)=0.1% MIS(clothes)=0.2%

지지도가 0.15%인 clothes→bread 인 규칙에 적용되는 최소 지지도는 mim(MIS(clothes), MIS(bread)) = 0.2이므로 이 규칙은 최소지지도를 만족하지 못하는 규칙이다. 그러나 같은 지지도 0.15%인 clothes→shoes의 경우, 이 규칙의 최소지지도는 min(MIS(clothes), MIS(shoes)) = 0.1 이므로 최소 지지도를 만족하는 규칙임을 알수 있다.

MSApriori 알고리즘은 각각의 데이터에 대한 빈도형태를 고려하여 데이터마다 다른 지지도인 MIS를 사용하여 연관규칙을 탐사하는 방법이며 이러한 방법으로 Apriori 알고리즘에서의 일괄적인 지지도 적용으로 생기는 문제점을 보완할 수 있다. 그러나 MSApriori 알고리즘은 데이터베이스에 존재하는 데이터 모두에게 MIS를 지정하기 위해서 모든 데이터들의 빈도 형태를 연관규칙의 탐사 단계 이전에 파악되어야 하는 문제가 발생한다. 따라서 데이터베이스를 구성하는 모든 항목에 대해서 빈도수를 파악하고 그에 맞는 MIS를 설정하기 위한 전처리기 과정이 추가되며 데이터베이스가 많은 항목들로 구성된 경우는 상당히 복잡해지게 되는 문제가 발생하게 된다.

4) RSAA 알고리즘

데이터 베이스에서는 상대적으로 드물게 나타나지만 고가의 의미를 가지고 있는 고부가가치 제품이 존재하며, 이러한 제품은 판매횟수에 비해 많은 이익을 제공할 수 있다. 실세계에서는 데이터베이스에서 차지하는 통계적인 비중

은 적지만 이윤 추구 등에 있어서 중요한 희소 데이터들이 존재한다. 여기서 의미있는 희소 데이터란 데이터베이스에서 발생 빈도수가 최소지지도를 만족하지 못하지만 데이터의 발생 빈도수 중 높은 비율로 특정 데이터들과 연관되어 나타나는 데이터이다.

그러나 Apriori와 같은 기존의 연관 규칙 탐사 알고리즘의 경우, 데이터베이스에 일괄적인 지지도를 적용하기 때문에 데이터베이스의 상이한 데이터의 빈도수의 형태를 고려하지 못하므로 의미 있는 희소 데이터를 탐사할 수 없는 경우가 발생한다. 그래서 RSAA(Relative Support Assocition Apriori)[32]는 데이터 사이의 상대적인 빈도수를 고려하여 연관성을 탐사할 수 있는 방법으로써 상대 지지도를 사용하여 희소 데이터들에 존재하는 연관규칙을 탐사하는 방법이다.

RSAA에서는 2개의 최소 지지도가 설정된다. 하나는 빈발항목 탐사 시 사용자가 지정하는 최소 지지도로써 이 값을 만족하는 항목과 만족하지 못하는 항목은 빈발 항목과 희소한 항목으로 구분된다. 다른 하나의 지지도는 희소한 데이터에 대하여 상대 지지도를 적용할 대상을 평가하는 최소 지지도이다. RSAA에서 사용되는 2종류의 지지도에 대한 정의는 다음과 같다. 1차 지지도은 빈발항목 탐사과정에 사용되기 위하여 사용자가 정의한 지지도의 임계값이고, 2차 지지도는 희소항목 탐사과정에 사용되기 위하여 사용자가 정의한 지지도의 임계값이다. 1차 지지도와 2차 지지도의 설정은 (1차 지지도 > 2차 지지도)를 만족하도록 설정해야 한다. 그렇지 않을 경우, 중복된 규칙이 생성되거나 희소항목에 대한 탐사를 하지 못하기 때문이다.

RSAA는 데이터 사이의 상대적인 빈도수를 고려하여 연관 규칙을 탐사할 수 있는 상대지지도를 사용한다. 상대지지도는 2차 지지도를 만족하는 희소

32) 하단심, "의미 있는 희소데이터를 포함한 연관 규칙 탐사기법", 전남대학교 대학원 석사 논문, 2001.

항목과 임의 다른 데이터 사이의 상대적인 지지도이며, 상대 지지도를 이용해 의미있는 희소한 데이터들을 탐사할 수 있다. 상대지지도 (Relative Support)는 데이터베이스가 데이터의 집합 I={i_1, i_2, i_3 … i_m}와 같이 구성되고 데이터 항목 i의 지지도가 sup(i)로 표현된다면, 데이터 사이의 상대적인 지지도를 의미하는 상대지지도 Rsup(i_1, i_2, i_3 … i_m)의 후보항목집합 i_1, i_2, i_3 … i_k에서의 정의는 아래와 같다.

$$\begin{aligned} Rsup(i_1, i_2, i_3 \cdots i_k) = \max(&sup(i_1, i_2, i_3 \cdots i_k)/sup(i_1) \\ &sup(i_1, i_2, i_3 \cdots i_k)/sup(i_2) \ldots \\ &sup(i_1, i_2, i_3 \cdots i_k)/sup(i_k)) \end{aligned}$$

상대지지도는 0≤Rsup≤1인 값으로써 데이터 항목을 구성하는 각각의 항목들이 후보 항목과 이루는 신뢰도 값들을 비교하여 이 중 가장 큰 값을 선택한다. 이는 후보 항목을 구성하는 각각의 i_1, i_2, i_3 … i_k들이 후보항목 {i_1, i_2, i_3 … i_k}에 대하여 얼마 만큼의 비중으로 함께 나타나는지를 나타내는 척도이며 사용자는 적절한 최소상대지지도(minimum relative support)를 입력하여 규칙을 탐사한다.

최소상대지지도 (Minimum Relative Support)는 사용자에 의해 정해진 상대지지도의 임계값으로 minRsup로 표기한다. 준빈발 항목집합은 2차 지지도를 만족하는 항목 집합에서 최소상대지지도를 만족하는 항목의 집합이다.

상대지지도는 1차 지지도는 비록 만족하지 못하지만 2차 지지도를 만족하는 희소한 데이터를 대상으로 적용된다. RSAA는 준 빈발항목 탐사 단계에서 상대 지지도를 계산하며, 계산된 상대지지도의 값이 최소상대지지도를 만족하면 준 빈발항목이 된다. 최소상대지지도 값이 높을수록 사용자는 동시에 나타나는 비율이 큰 항목을 선택함을 의미한다.

상대지지도는 준빈발항목 집합 탐사 단계에서의 신뢰도 사용과 같다. 기존의 Apriori 알고리즘의 빈발항목 구성단계는 하나의 지지도를 이용하여 연관규칙의 생성에 쓰일 항목들을 추려낸다. Apriori 알고리즘은 지지도를 만족하는

빈발항목집합에서 규칙을 생성한후, 신뢰도를 적용하여 규칙의 타당함을 검증하는 방식으로 진행된다. 그러나 RSAA는 빈발항목 뿐만 아니라 희소항목에 대하여 신뢰도의 의미를 갖는 상대지지도를 적용하여 특정 데이터와 높은 비율로 동시에 나타나는 희소 데이터를 탐사한다. 또한 데이터의 계층이 존재하는 경우의 연관 규칙 탐사에서도 RSAA를 적용할 수 있는데, 상위 단계보다 상대적으로 희소한 하위 단계의 데이터 항목에 관하여 연관규칙 탐사가 필요한 경우, RSAA를 이용하여 하위 단계의 연관규칙이나 상위단계이지만 다른 데이터보다 희소한 성격을 갖는 데이터를 포함한 연관규칙을 탐사 할수 있다.

RSAA의 후보 항목 생성 방법은 희소 데이터를 포함한 후보항목을 구성할 수 있어야 한다. RSAA에서 후보 항목은 2부분으로 나누어 구성이 된다.

```
Deviding between Large Itemset and Rare Data Itemset Algorithm
I={데이터베이이스의 모든 항목}
support1 = 1차 지지도
support2 = 2차 지지도
  for each item i ∈ I do
       if i.support ≥ support1 then
            i ∈ Ck
       else
            if i.support ≥ support2 then
               i ∈ NCk
       end
end
```

<알고리즘 4> 빈발 항목데이터와 희소 데이터의 분리

RSAA에서의 후보항목 구성은 1차 지지도를 만족하는 빈발항목과 1차 지지도를 만족하지 않지만 2차 지지도를 만족하는 항목들로 나누어서 구성이 된다. 1차 지지도를 만족하는 원소들은 기존의 Apriori의 방식과 동일하다. RSAA에서의 후보 항목 구성은 2차 지지도를 만족하는 항목들을 대상으로 구성된다.

RSAA에서의 희소데이터와 빈발항목 데이터를 분할 하는 방법은 <알고리즘 4>와 같다.

<알고리즘4>은 데이터베이스의 모든 데이터에 대해 지지도를 카운트하여 1차 지지도를 만족하는 빈발항목에 대한 1-후보항목집합인 C_1과 1 차지지도를 만족하지 못하지만 2차 지지도를 만족하는 희소 데이터 항목 집합의 후보항목 집합 NC_1으로 각각 분리 되어 생성한다.

분리된 희소 데이터 항목에 대한 후보항목집합의 구성은 NC_k(Not Frequent Candidateset), NLC_k(Not frequent and join Large item generated Candidateset)의 2 그룹으로 나누어 각각 생성된다. 2-후보항목집합은 NC_2와 NLC_2로써 각각은 다른 조인과정을 거친다. NC_2는 NL_1(Not Large itemset)의 조인으로 생성되며 NLC_2는 1-빈발 항목집합 L_1과 준빈발항목집합 NL_1의 조인으로 생성한다. NC_k와 NLC_k는 각각의 조인을 통해 생성된다.

NL_k는 NC_k 에 대하여 2차 지지도와 상대지지도를 평가하여 생성되고, NLL_k는 NLC_k 에 대하여 생성된다. 준빈발 항목집합은 NL_k와 NLL_k를 의미한다. <알고리즘 5>는 RSAA의 후보 항목 생성 알고리즘이다.

```
Creating Candidate Itemset for Rare Data Algorithm
If (k=2) then
        insert into NC2
                select p.item1 ,    q.item1   from   NL1 p, NL1 q
```

```
        insert into NLC2
                select p.item1,    q.item1 from NL1 p, L1 q
else
        insert into NCk
                select p.item1 , p.item2 , ... p.itemk-1, q.itemk-1
                from NLk-1 p, NLk-1 q
                where p.item1 = q.item1, p.item2 = q.item2, ...,
                p.item k-1 = q.itemk-1, p.itemk-1 < q.itemk-1
        insert into NLCk
                select p.item1 , p.item2, ... ,p.item k-1 , q.item1
                from NLL k-1 p, NLL k-1 q
                where p.item1=q.item1.p.item2= qitem2, ... ,
                p.item k-1 = q.item k-1,    p.item k-1 < q.item k-1
```

<알고리즘 5> 희소 데이터에 대한 후보항목 구성

2. 기존 알고리즘의 분석

Apriori에서는 항목이 빈발하게 나타나는 패턴을 연구하였다. 그러나 이러한 연구에서 지지도가 높으면 필요한 정보가 삭제되고, 지지도가 낮으면 필요 이상의 정보가 검출되는 단점을 보안하기 위해서 MSApriori에서는 ß라는 값을 각각의 데이터의 빈도수에 곱한후, 그 값을 Ls와 비교하여 데이터들의 MIS를 구하게 된다. 이러한 방법으로 Apriori 알고리즘에서의 일괄적인 지지도 적용으로 생기는 문제점을 보완 할 수 있으나, MSApriori 알고리즘은 데이터베이스에 존재하는 데이터 모두에게 MIS를 지정하기 위해서 모든 데이터들의 빈도 형태를 연관규칙의 탐사단계 이전에 파악 되어야 하는 문제가 발생한다. 따라서 데이터베이스를 구성하는 모든 항목에 대해서 빈도수를 파악하고 그에 맞는 MIS를 설정하기 위한 전처리 과정이 추가 되며 데이터베이스가 많은 항목들로 구성된 경우는 상당히 복잡하게 되는 문제가 발생한다.

RSAA에서는 희소한 데이터가 특정 데이터에 대하여 빈발하게 나타나는 경우에 대해서 연관규칙을 탐사하였다. 희소 데이터항목을 탐사 대상으로 하는 이유는 중요한 데이터 항목이 존재 할 것이라는 가정이 기반이 된다. 그러나 희소한 데이터들의 중요성 정도를 고려하지 않으므로 필요 이상의 준빈발 항목를 만듦으로써 필요하지 않는 항목도 탐색하게 되고, 대용량의 데이터에 있어서 처리 속도가 늦어지게 된다. 그러므로 데이터항목의 중요도를 고려하여 2차 빈발항목의 개수를 줄여줌으로 해서 보다 빠른 처리 속도와 보다 의미 있는 정보를 추출할 수 있다.

데이터항목의 중요성 정도는 이미 탐사전에 탐색자에 의해 인식을 할수 있다. 할인매장의 경우, 생필품과 전자제품 항목을 같이 비교한다면 누구나 매출이익이 전자제품이 높다는 것을 알수 있다. 그러므로 이미 항목간의 중요성이 인식될 수 있다. 또한 중요도의 기준을 전에 탐색했던 자료를 근거로 하여

정할 수도 있다. 데이터 항목들간의 중요도를 고려하여 중요하지 않은 항목은 버림으로써 보다 중요하고 의미있는 연관규칙을 보다 빠르게 탐사할 수 있다.

Ⅳ. 새로운 연관규칙 탐사 알고리즘

이 장에서는 기존의 RSAA알고리즘을 보완하여 상대적으로 연관성이 높은 희소한 데이터항목 중에서도 중요도가 높은 데이터항목 만을 대상으로 마이닝을 할수 있는 새로운 알고리즘을 제안한다.

1. 중요지지도의 정의

실세계의 상황에서 데이터항목간의 가치는 다르다. 예를 들면 장바구니 분석에 있어서 휴지와 텔레비전의 매출이익은 다르다. 휴지를 10개 팔아도 텔레비전을 1개 파는 매출이익과는 비교 할 수 없다. 텔레비전을 판매하는 것이 당연히 높다. 그러므로 판매사원은 휴지와 텔레비전을 판다고 하면 텔레비전에 더욱 신경을 쓰게 될 것이다. 또한 매출이익의 차이가 적은 경우에는 많이 팔리는 항목에 대해서 판매촉진전략을 세울 것이다. 예를 들면, A라는 항목은 10원의 판매이익이 생기고, B라는 항목은 20원의 판매이익이 생긴다. 그런데, B보다는 A를 소비자가 더 선호 한다면, B를 한 번 파는 것과 A를 두 번 파는 것 중 A가 더 쉽다고 할 수 있다. 그러므로 판매사원은 B보다는 A의 판매촉진에 더욱 관심을 가지게 된다. 위의 경우에서 항목간의 속성의 가치는 빈도수 뿐만 아니라, 그 항목이 갖고 있는 속성의 가치에 의해서도 결정이 된다.

데이터 항목의 중요순위(Order of Importance)란 데이터항목의 매출이익, 비용, 선호도, 심리적 가치, 그 전 탐사의 결과, 속성의 가치 등에 의해서 순차적으로 결정된 서열이다. 중요가중치(Importance Weight)란 데이터 항목의 중요순위를 기초로 데이터항목의 중요도를 가름할 수 있는 척도이다. 중요가

중치를 아래와 같이 정의한다.

<정의1> 중요가중치(Importance Weight)

데이터베이스의 트랜잭션 데이터는 데이터항목의 집합 $I = \{i_1, i_2, i_3 \cdots i_m\}$ 를 포함하고 각 항목의 중요순위는 $i_1 < i_2 < i_3 < \cdots < i_m$ 일 때 데이터 항목 i_k의 중요순위값은 k라고 하자. 이 때 i_k의 중요가중치 w_k는 다음과 같다.

$$w_k = \frac{k}{\sum_{a=1}^{m} i_a}$$

□

또한, 임의의 연관규칙을 구성하는 데이터항목 집합의 전체적인 중요성 정도를 나타낼 수 있는 중요지지도(Weight Support)는 다음과 같이 정의된다.

<정의2> 중요지지도(Weight Support)

데이터항목 집합 $\{i_1, i_2, i_3 \cdots i_k\}$의 중요지지도 $Wsup(i_1, i_2, i_3 \cdots i_k)$는 아래와 같이 정의된다.

$$Wsup(i_1, i_2, i_3 \cdots i_k) = \min(w_1, w_2, w_3 \cdots w_k) \quad \square$$

2. WRSAA 알고리즘

RSAA 알고리즘이 희소데이터들 사이의 상대적인 연관성만을 고려하는 반면, 제안하고자 하는 새로운 알고리즘은 희소데이터들 사이의 상대적인 연관

성뿐만 아니라 데이터항목의 가중치를 기반으로 희소데이터의 중요성 정도를 고려하여 연관규칙탐사를 한다. 이러한 의미에서 제안하는 새로운 알고리즘을 WRSAA(Weighted Relative Support Assocition Apriori)라고 부르기로 한다.

WRSAA알고리즘은 2차 지지도와 상대지지도 그리고 중요지지도라는 척도를 이용하여 비록 데이터베이스에서는 희소하지만 높은 비율로 동시에 발생하는 중요한 데이터항목들사이의 연관성을 추출해 낼 수 있다.

WRSAA의 알고리즘은 <알고리즘6>와 같다. <알고리즘6>에서 사용되는 자료 구조와 함수의 역할은 그림4-1와 같다.

minWsup	최소 중요 지지도
support1	1차 지지도
support2	2차 지지도
minRsup	최소 상대 지지도
Wsup	데이터 항목의 중요 지지도
Rsup	데이터 항목의 상대 지지도
k-itemset	K개의 아이템으로 구성된 집합
NC_k, NLC_k	희소항목에 대한 k-후보항목집합
NL_k ,NLL_k	k-준빈발항목집합
rsaa-gen	준빈발항목집합에 대한 후보항목 생성 함수
subset	트랜잭션에 후보항목이 존재하는지 검사하는 함수

<그림 4-1> WRSAA 에서 사용되는 자료 구조와 함수들

WRSAA Algorithm

D: Database

$I = \{i_1, i_2, \ldots, i_k\}$

for all $(i_k \in I)$ /*<블록1>*/

if (i_k.support $\geq$ $support_1$) then $i_k \in L_1$

else if (i_k.support $\geq$ $support_2$) then $i_k \in NL_1$

end

each item $\in$ L_1 /*<블록2>*/

do Apriori Algorithm

end

if (k=2) /*<블록3>*/

NC_2 = rsaa-gen (NL_1, NL_1);

NLC_2= rsaa-gen (NL_1, L_1);

end

for (k=3; $NL_{k-1} \neq \emptyset$ or $NLL_{k-1} \neq \emptyset$;k++) do /*<블록4>*/

NC_k= rss-gen(NL_{k-1}, NL_{k-1}); /*<서브블록1>*/

NLC_k=rss-gen(NLL_{k-1}, NLL_{k-1});

```
for all transaction t ∈ D do                    /*<서브블록2>*/
    NC_t = subset(NC_k, t);
    NLC_t = subset (NLC_k, t);

    for all candidates nc ∈ NC_t do             /*<서브블록3>*/
        nc.count++;
    end

    for all candidates nlc ∈ NLC_t do           /*<서브블록4>*/
        nlc.count++;
    end

    if nc.count ≥ support_2 then                /*<서브블록5>*/
     each item i_k in nc
        nc.Rsup = max( sup(i_1, i_2, i_3 ··· i_k)/sup(i_1),
                       sup (i_1, i_2, i_3 ··· i_k)/sup(i_2),...
                       sup (i_1, i_2, i_3 ··· i_k)/sup(i_k))
        nc.Wsup = min(w_1, w_2, w_3 ··· w_k)

if ((nc.Rsup ≥ minRsup) and (nc.Wsup ≥ minWsup)) then
NL_k={nc∈ NC_t | (ncRsup ≥ minRsup) and (nc.Wsup ≥
                                          minWsup) }
    end

    if nlc.count ≥ support_2 then               /*<서브블록6>*/
```

```
                each item ik in nlc
                 nlc.Rsup= max(sup(i1, i2, i3 ···  ik)/sup(i1),
                                sup (i1, i2, i3 ···  ik)/sup(i2), ...,
                               sup (i1, i2, i3 ···  ik)/sup(ik))
                          nlc.Wsup = min(w1, w2, w3 ···  wk)
            end
    if (nlc.Rsup ≥  minRsup) and (nlc.Wsup ≥ minWsup) then
              NLLk ={nlc∈NLCt | (nlc.Rsup ≥minRsup) and
                                        (nlc.Wsup ≥ minWsup) }
            end
        end
end
Answer = Uk NLk , Uk NLLk     /*<블록5>*/
  end
```

<알고리즘 6> WRSAA 알고리즘

<알고리즘6>에서 <블록1> 부분은 데이터베이스 D에 있는 각 데이터 항목들의 지지도를 1차 지지도와 2차지지도와 비교하여 1-빈발항목과 1-준빈발항목을 구분해내는 과정이다. <블록2> 부분에서는 1-빈발항목들에 대하여 Apriori 알고리즘을 적용하여 전체적인 빈발항목집합을 생성한다.

<블록3> 부분에서는 2-준후보항목집합을 생성하는 과정이다. <블록4> 부분에서는 2-준후보항목집합을 시발점으로 하여 전체적인 준빈발항목집합을 생성하는 과정이다. 여기에서 생성되는 준빈발항목집합들은 비록 1차지지도는 만족하지 못하지만 2차지지도, 최소상대지지도, 최소중요지지도를 만족한다.

<블록4> 안에 있는 <서브블록1>에서는 k-1-준빈발항목집합으로부터 k-준후보항목집합을 생성한다. <서브블록2>에서는 <서브블록1>에서 생성된 k-준후보항목집합안의 각 데이터항목집합들의 지지도를 계산하기 위하여 데이터베이스 D의 트랜잭션들을 검토하는 과정이다. 이러한 검토과정 후에 <서브블록3>과 <서브블록4>에서는 k-준후보항목집합안의 각데이터항목집합들의 지지도를 계산한다. <서브블록5>와 <서브블록6>에서는 k-준후보항목집합 중 최소상대지지도와 최소중요지지도를 만족하는 것들만을 추출하여 k-준빈발항목을 생성하는 과정이다. <블록4>의 for루프를 통하여 더 이상의 k-준빌발항목집합이 생성되지 않으면 for 루프를 빠져나와서 <블록5> 부분에서 이제까지 생성된 준빈발항목들을 유니온(union)하여 최종적으로 전체적인 준빈발항목집합 Answer를 생성한다.

3. 중요지지도를 고려한 연관규칙 탐사의 예

여기에서는 WRSAA알고리즘이 연관성이 높은 희소한 데이터 중에서도 중요한 데이터항목만을 대상으로 마이닝을 수행하는 예를 고찰한다.

<표4-1>은 트랜잭션에 포함된 데이터항목들(1, 2, 3, 4, 5, 6, 7)을 나타내고 있고 데이터항목들의 중요순위는 1 < 2 < 3 < 4 < 5 < 6 < 7 라고 가정한다. 데이터마이닝을 위하여 1차지지도는 40%, 2차지지도는 20%, 최소상대지지도는 0.7이며 최소중요지지도는 0.1로 설정한다.

<표 4-1> 중요지지도를 고려한 연관규칙의 예

TID	항목	TID	항목
1	2,3	6	4,6
2	3,4	7	3,4,6,7
3	2,3,4,5,6,7	8	1,2
4	6,7	9	3,4,5,
5	3,4,5,6	10	1,2

트랜잭션을 구성하는 데이터항목들에 대한 중요지지도는 <표 4-2> 와 같다.

<표4-2>트랜잭션의 구성요소에 대한 중요지지도

항목	지지도	중요 지지도
{1}	2	(1/28)=0.04
{2}	4	(2/28)=0.07
{3}	6	(3/28)=0.1
{4}	6	(4/28)=0.14

{5}	3	(5/28)=0.18
{6}	5	(6/28)=0.2
{7}	3	(7/28)=0.25

<표4-2>에서 NL_1 = {1, 5, 7}은 준빈발항목집합으로 1차 지지도를 만족하지 못하지만 2차 지지도를 만족하는 데이터이다. L_1 = {2, 3, 4, 6}은 빈발항목집합이다. 빈발항목집합은 Apriori 알고리즘에 의해서 처리된다. 다음은 탐색된 L_1과 NL_1에 대하여 NL_1*L_1 와 NL_1*NL_1 에 대하여 2-후보항목 NC_2, NLC_2를 생성하다. <표4-3>에서 후보 항목에 대한 상대지지도가 0.7이상인 항목 집합은 {{1,2},{5,3},{5,4}}이다. 이중에서 최소 중요지지도 0.1 이상의 항목은 {{5,3},{5,4}}가 된다. {1,2}은 중요지지도가 0.04이므로 상대지지도가 높지만 전지한다. 그러므로 3-후보항목 NLC_3 은 {5,3,4}가 된다. 이상에서 알수 있는 것과 같이 WRSAA는 상대지지도가 높은 희소한 데이터중에서도 중요지지도가 높은 데이터 항목만을 대상으로 준빈발항목집합을 생성함을 알수 있다.

<표 4-3 >NC_2와 NLC_2 의 구성요소에 대한 지지도

NC_2	지지도	상대지지도	중요지지도
{1,2}	2	1	0.04
{1,3}	0	-	-
{1,4}	0	-	-
{1,6}	0	-	-
{5,2}	1	0.3	-
{5,3}	3	1	0.1
{5,4}	3	1	0.14

{5,6}	3	0.6	-
{7,2}	1	0.3	-
{7,3}	2	0.6	-
{7,4}	2	0.6	-
{7,6}	2	0.6	-
NLC_2	지지도	상대지지도	중요지지도
{1,5}	0	-	-
{1,7}	0	-	-
{5,7}	1	0.6	-

V. 성능 평가 및 분석

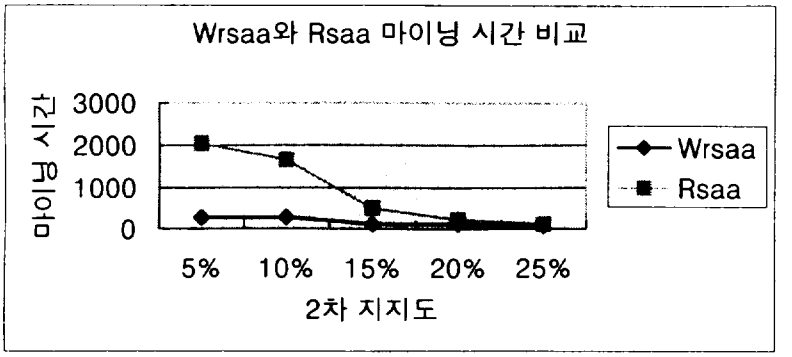
이 장에서는 기존의 RSAA보다 WRSAA가 중요한 희소항목을 대상으로 한 연관규칙 탐사 시에는 보다 효율적임을 보이고자 한다. WRSS와 RSAA를 비교하기 위한 기준은 3가지로 설정하였다. 첫째는 모든 빈발항목집합들을 생성하는 시간, 둘째는 준후보항목 개수, 셋째는 준빈발항목 개수이다. 후보항목과 빈발항목집합은 WRSAA와 RSAA 모두 Apriori 알고리즘을 이용하기 때문에 평가 대상에서 제외 하였다.

WRSAA와 RSAA 알고리즘의 구현은 C언어를 이용하였으며, 구현된 알고리즘은 펜티엄Ⅳ의 Windows Me 환경에서 실험하였다. 임의의 트랜잭션는 10,000개이며, 항목의 개수는 55개, 트랜잭션당 최대 항목의 개수는 14개로 설정하였다. 트랜잭션을 구성하는 Data set은 Random함수을 이용하여 생성하였고, 연관성이 있는 희소데이터 비율은 전체 데이터의 25% 수준으로 설정하였다. Dataset에 대하여 상대지지도는 0.5, 중요지지도는 0.5, 1차지지도는 70%로 설정하고 RSAA와 WRSAA를 비교하였다.

모든 준빈발 항목을 생성하는 시간은 <표 5-1>에서 처럼 RSAA보다 WRSAA가 짧은 것으로 나타났다. 이는 WRSAA의 경우 중요지지도가 0.5 미만인 데이터는 마이닝에서 제외함으로써 처리해야 할 준후보항목 및 준빈발 항목의 수가 줄어 들었기 때문이다.

<표 5-1> Wrsaa와Rsaa 마이닝 시간 비교

2차 지지도 / 기법	5%	10%	15%	20%	25%
Wrsaa	270	280	110	110	60
Rsaa	2040	1650	500	220	110

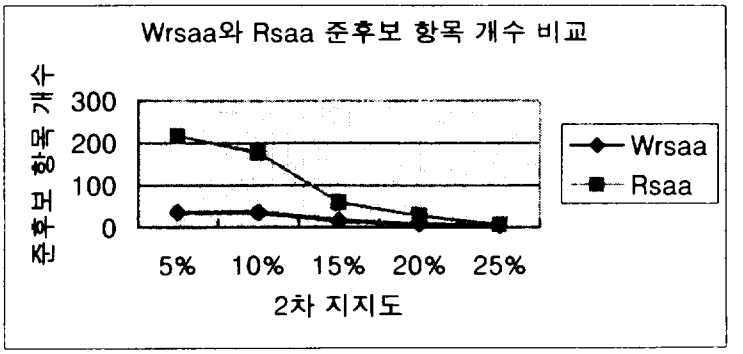


<그림 5-1> WRSS와 RSAA 마이닝 시간 비교

<표 5-2>에서 준후보 항목 개수를 비교해 본 결과 2차 지지도의 변화에 대하여 준후보 항목 개수가 RSAA에 비하면 WRSAA에서는 준후보 항목이 적다. 왜냐하면 중요하지 않는 항목을 제외 되었기 때문이다. 준후보 항목의 개수가 줄어들면 탐색의 시간이 줄어들고, 중요한 항목만 검색됨으로써 결과에 대한 신뢰도가 높아진다.

<표 5-2> 준후보 항목 개수 비교

2차 지지도 / 기법	5%	10%	15%	20%	25%
Wrsaa	36	35	16	7	3
Rsaa	216	177	59	28	6

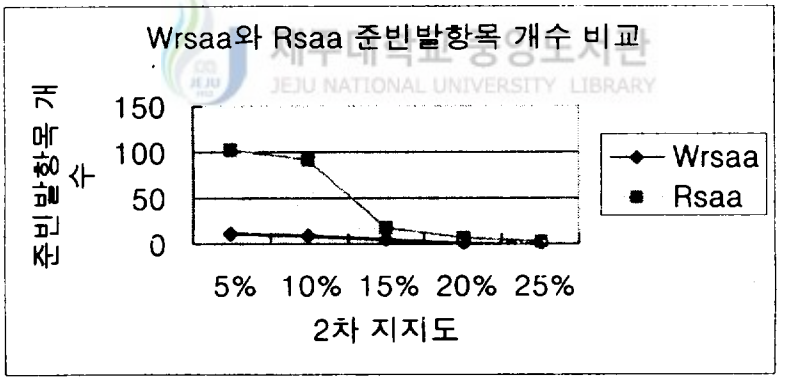


<그림 5-2> 준후보 항목 개수 비교

다음은 <표 5-3>에서 준빈발항목개수를 비교해본 결과 중요지지도가 낮은 항목은 제외되었기 때문에 준빈발항목 개수의 비율이 RSAA에 비해 WRSAA가 적어진다. 그러므로 준빈발 항목 개수가 줄어듦으로써 탐색의 효율이 높아진다.

<표 5-3> 준빈발항목 개수 비교

2차지지도 / 기법	5%	10%	15%	20%	25%
Wrsaa	10	8	4	1	0
Rsaa	102	92	17	7	3



<그림 5-3>준빈발항목 개수 비교

<표5-4>에서는 WRSAA에 대하여 상대지지도를 0.5, 1차지지도는 70%, 2차 지지도는 5%로 설정하였을 때, 중요지지도 변화에 따른 후보 및 빈발항목 개수의 변화, 데이터 마이닝시간의 변화를 나타내고 있다. 최소중요지지도가 0일때는 중요지지도를 고려하지 않는 상황이 되기 때문에 <표5-2>,<표5-3>에서 알수 있드시 RSAA 알고리즘과 동일한 성능을 나타내게 된다. 따라서,

WRSAA알고리즘은 RSAA알고리즘의 기능을 수용할 수 있는 보다 일반화된 알고리즘이라고 할 수 있다.

최소중요지지도가 증가하면 준후보항목과 준빈발항목의 개수가 감소된다. 최소중요지지도가 높아지면 항목들의 중요지지도가 높은 항목에 대해서만 탐색이 이루어진다. 그러므로 마이닝 시간은 짧아지고, 후보 항목들의 개수가 적어지므로 탐색의 효율이 높아진다.

<표 5-4> 최소중요지지도 변화에 따른 후보 및 빈발 항목 개수의 변화

중요지지도	0.0	0.1	0.2	0.3	0.4
준후보 항목	216	203	147	79	60
준빈발 항목	102	101	64	27	20
마이닝 시간	1980	1870	1320	660	490
중요지지도	0.5	0.6	0.7	0.8	0.9
준후보 항목	36	29	20	17	14
준빈발 항목	10	7	4	3	1
마이닝 시간	280	220	110	110	60

Ⅵ. 결론

최근 대용량의 데이터베이스의 구축이 보편화되면서 단순한 데이터의 저장뿐만 아니라 저장된 데이터 속에 숨겨진 지식 추출의 필요성이 대두되었다. 데이터 마이닝은 이러한 필요성에 의해 급부상 되고 있으며 기업, 연구소, 전자 상거래, WWW과 같은 다양한 환경에서 응용되고 있다.

데이터 마이닝의 많은 기법 중 가장 활발히 연구되고 있는 분야는 연관 규칙 탐사 분야이다. 연관 규칙은 데이터베이스에서 빈발하게 동시에 나타나는 강한 연관성을 지닌 항목들을 탐사하여 규칙의 형태로 표현한 것으로 지지도라는 척도를 이용하여 빈발한 항목들을 탐사한다. 기존의 연관규칙의 탐사 방법의 경우, 데이터베이스에 대해 일괄적인 지지도를 적용하므로 데이터베이스에 존재하는 모든 데이터항목들은 유사한 빈도수를 가지고 있는 것으로 가정한다. 즉 지지도를 만족하지 못하는 모든 항목들은 연관규칙 탐사 과정에서 제거되게 된다.

그러나 실세계에서는 데이터항목의 성격에 따라 데이터베이스상에 상대적으로 빈발하게 나타나는 데이터항목도 존재하고, 반대로 상대적으로 희소하게 나타나는 항목도 존재한다. 또한, 희소하게 나타나는 데이터항목들 중에서도 강한 연관성을 가지는 규칙이 존재할 수 있고 이러한 데이터항목들은 고가의 제품과 같이 이윤 마진 폭이 큰 중요한 데이터일 수 있다.

기존의 대부분의 연관규칙탐사 알고리즘들은 빈발하게 나타나는 데이터항목들만을 대상으로 탐사작업을 수행하였고, 비록 빈발하게 나타나지 않는 데이터항목 들을 대상으로 연관규칙을 탐사하는 알고리즘들도 있으나 이러한 알고리즘들은 데이터항목사이의 연관성만을 고려하기 때문에 중요성이 있는 희소데이터들 사이의 연관성을 탐사하려는 원래의 목적에 충실한 연관규칙을 탐사할 수 없었다.

본 논문에서는 데이터항목들에 중요가중치를 부여하여 중요가중치를 기반으로 데이터항목들의 중요성 정도를 측정하고 어느 정도의 중요지지도를 만족하는 연관규칙을 탐사하는 알고리즘을 제안하였다. 본 논문에서 제안한 WRSAA 연관규칙탐사 알고리즘을 활용함으로써 어느 정도의 중요성을 가지면서도 희소하게 나타나는 데이터항목들에 대한 연관성 탐사를 보다 빠르게 처리할 수 있다. 또한, 최소중요지지도를 조절하여 기존의 RSAA알고리즘의 기능을 수용할 수 있게 할 수 있으므로 WRSAA는 기존의 RSAA 알고리즘을 일반화시켰다고 볼 수 있다.

기업활동의 결과로 생성된 데이터들은 갈수록 많아져서 데이터 과잉문제가 발생하는 현실을 감안할 때 보다 중요한 데이터들만을 대상으로 연관규칙을 탐사를 실행하는 WRSAA알고리즘은 기업 데이터 분석에 보다 유용하게 적용될 수 있을 것으로 기대한다.

향후 연구 방향으로는 알고리즘의 효율적인 개선과 현장에 적용, 분석함으로써 실제 사례에 대한 폭 넓은 응용연구가 필요하다.

<참고 문헌>

국내문헌

김재경, 이건창, 정남호, 권순재, 조윤호, "클레멘타인 데이터마이닝 솔루션을 이용한 웹 로그 분석", Information Systems Review, Vol. 4, No.1, 2002.3, pp.47~60.

김정자, 이도현, "데이터 마이닝 기술 및 연구동향", 정보과학학회지, 제 16권 제 9호, 1998.9, pp.6~14.

남도원, 김성민, 이동하, 오재훈, 김성훈, 이전영, "한시적 연관규칙에서의 부분 구간 탐사", 한국정보과학회 데이터베이스 학술대회, 1999.

박원환, "수량 연관규칙을 위한 데이터 마이닝 알고리즘에 관한 연구", 순천향 대학교 대학원 석사 논문, 2001.

박정호, "Apriori 알고리즘 연관 규칙마이닝 기법을 이용한 정보검색", 고려대학교 대학원 석사 논문, 1999.

박종수, 유원경, 홍기형, "연관 규칙 탐사와 그 응용", 정보과학 학회지, 제16권,1998.

안효성, "연관 규칙을 활용한 데이터베이스 지식 탐색 도구 구현에 관한 연구", 국민대학교 대학원 석사 논문, 1999.

이도현, "데이터 마이닝을 이용한 CRM" 정보과학회지, 제18권 11호, 2000, pp.4~11.

이재규, 최형림, 김현수, 이경전, "전자상거래원론", 법영사, 1999.

이정원, 김호숙, 최지영, 김현희, 용환승, 이상호, 박승수, "데이터마이닝 알고리즘의 분류 및 분석", 정보과학회 논문지, 데이터베이스 제 28권 제 3호, 2001.9, pp.279~299.

장남식, 홍석완, 장재호, "데이터 마이닝" 대청미디어, 2000, pp.19~49.

정희택, "고객 관계 관리를 위한 워크플로우 시스템의 통합", 정보과학회지, 제 18권 11호, 2000, pp.22~28.

조재희, 박성진, "데이터 웨어하우징과 OLAP", 대청, 1996.

최영희, 장수민, 유재수, 오재철, "수량적 연관규칙탐사를 위한 효율적인 고빈도 항목열 생성기법", 한국정보처리학회논문지, 제6권 제10호, 1999, pp.2597-2607.

하단심, "의미 있는 희소데이터를 포함한 연관 규칙 탐사기법", 전남대학교 대학원 석사 논문, 2001.

황인수, "고객관계관리에서 신경망을 이용한 제품-고객군의 형성에 관한 연구", 경영정보학연구, 제 11권 제 4호 , 2001, pp.27~41.

황정희, 신예호, 류근호, "트리거와 점진적 갱신기법을 이용한 연관규칙탐사의 능동적 후보 항목 관리 모델", 정보과학회논문지, 데이터베이스 제 29권 제1호, 2002.2, pp.1~13.

황현숙, 어윤양, "연관 마이닝과 고객 선호도 기반의 인터넷 상품 검색 시스템 설계 및 구현" 경영정보학 연구 제12권 제1호, 2002.3, pp.1~16.

외국문헌

Bing Liu, Wynne Hsu, Yiming Ma, "Mining Association Rules with Multiple Minimum Supports", Proceedings of the ACM SIGKDD(KDD-99), 1999.

Chang, G, Healey, M. J., McHugh, J. A. M., and wang, J. T. L., "Mining the World Wid Web: An Information Search Approach", Kluwer Academic publishers, 2001.

Fayyad, U.M., G. Piatetsky-Shapiro, and P.Smyth, "From Data Mining to Knowledge Discovery," In Advances in Knowledge Discovery and Data Mining, Fayyad U.M, G.Piatetsky- Shapiro,P.Smyth and R.Uthurusamy, AAAI Press/Mit Press, CA., 1996, pp.1~34.

Heikki Mannila, "Methods and Problems in Data Mining", Proceeding of International conference on Database Theory", 1997.

Jiawei Han and Micheline Kamber, "Data Minig: Concepts and

Techniques", Morgan Kaufmann Pubishers, 2001.

Michael J. A Berry, and Gorden Linoff, "Data Mining Techniques: For Marketing, Sales, and Customer Support", John Wiley & Sons, Inc, 1997.

Pieter Adriaans, Dolf Zantige, "Data Mining", Addison Wesley, 1996.

R. Agrawal and R. Srikant, "Fast algorithms for Mining Association Rules", Proc. VLDB, 1994.

R. Agrawal and R. Srikant, "Mining Sequential Patterns", Proc. ICDE, March 1995.

R. Agrawal, Tomasz Imielinski, Arun Swami, "Mining Association Rules between Sets of Items in Large Database". Proccding of ACM SIGMOD, 1993.

Witten, I.H., and Frank, E., "DataMining: Practical Machine Learning Tools and Techniques with Java Implementations", Morgan Kaufmann Pubishers, 2000.

ABSTRACT

Study on Algorithm of Discovering Association Rules by applying Weight Support

Byung-ung Hwang
Department of Management Information Systems
Graduate School of Business Administration,
Cheju Natonal University
Supervised By Professor Keun-uyung Kim

Recently, as the large-scale database building has been generalized, the Data Mining area, which analyze stored data and find helpful knowledge of existing but unexposed in database, has been spotlighted as a new strategic technology for companies' marketing, electronic commerce and etc. Especially, it is helpful for management decision making by performing the role of reassuring the existing experiential knowledge retained by companies in corporate management and at the same time by offering new information and knowledge unrecognized till now.

The investigation technology for association rule in the area of Data Mining has been studied most actively, and applied to marketing, business management, and decision making of companies.

This research suggested the investigation algorithm for association rule that makes possible faster detection for more helpful information by considering the relative frequency and importance of database item in the

mass storage database at the same time.

The algorithm suggested in this research found to be better than existing one as result of simulation test.