데이터의 상대 지지도를 이용한 다단계 연관 규칙 탐사 기법

○
하 단 심, 황 부 현
전남대학교 전산학과
E-mail : {dsha, bhhwang}@sunny.chonnam.ac.kr

Discovery of Multiple-Level Association Rules using Relative Support of Data

Danshim Ha, Buhyun Hwang
Dept. of Computer Science, Chonnam National University

요 약

데이터는 다양한 빈도 형태와 속성을 가지고 있으며 데이터의 연관 규칙 탐사 시 이러한 데이터의 빈도수를 고려할 수 있는 방법이 필요하다. 그러나 기존의 연관 규칙 탐사 알고리즘은 지지도와 신뢰도만을 가지고 데이터의 연관성을 발견하며 데이터들의 발생 빈도는 고려하지 않는다.

본 논문에서는 하위 단계의 데이터나 동일한 단계지만 상대적으로 발생 빈도가 적은 데이터들의 연관 규칙을 탐사할 수 있는 방법을 제안한다. 제안하는 방법은 데이터의 상대 지지도를 이용한 다단계 연관 규칙 탐사 기법을 수행함으로써 데이터의 발생 빈도를 고려한 연관 규칙을 탐사할 수 있다. 그리고 탐사된 연관 규칙은 마케팅 분야 등의 여러 응용에서 유용하게 이용될 수 있다.

1 서 론

대용량의 데이터 수집과 처리가 용이해지면서 데이터에 존재하지만 분석되지 않은 지식의 탐사에 대한 연구가 활발히 진행되고 있다. 데이터 마이닝은 대용량의 데이터베이스에 존재하는 데이터들을 분석하여 존재하지만 드러나지 않은 유용한 지식을 발견해내는 과정을 의미한다[1,2]. 데이터 마이닝을 통하여 연관 규칙, 클러스터링, 분류 등의 지식을 얻을 수 있으며 얻어진 지식들은 장바구니 분석(market basket analysis), 비즈니스 관리 등의 응용에 이용 될 수 있다.

본 논문에서는 계층이 존재하고 발생 빈도수가 적은 희소 데이터를 대상으로 연관 규칙을 탐사할 수 있는 방법을 제안한다. 제안하는 방법은 데이터들간의 상대적인 발생 빈도수를 고려하는 척도인 상대 지지도를 이용하여 계층이 존재하는 데이터들에 대하여 상대적으로 적은 빈도수를 갖지만 높은 비율로 동시에 나타나는 데이터들을 탐사할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 연관 규칙의 정의와 탐사 기법, 다단계 연관 규칙 탐사 기법에 대해 소개하고 3장에서는 제안하는 다단계 연관 규칙 탐사 기법을 기술한다. 끝으로 4장에서 향후 연구 방향과 결론을 기술한다.

※ 이 논문은 한국과학재단 1999년도 특정기초연구비 (1999-2-303-006-3) 지원에 의하여 연구되었음.

2 관련 연구

2.1 연관 규칙의 정의

연관 규칙은 한 항목 그룹과 다른 항목 그룹 사이에 존재하는 연관성을 규칙의 형태로 표현한 것이다[3]. 연관 규칙 탐사는 사용자에 의해 적절하게 입력된 지지도(support), 신뢰도(confidence)라는 척도를 이용하여 데이터 상호간의 연관성을 파악할 수 있다.

두 항목 집합 X, Y가 존재하고 X와 Y 사이에는 어떠한 공통된 데이터가 존재하지 않을 경우($X \cap Y = \emptyset$), 이 두 항목 집합간에 연관성이 존재한다면 그 규칙은 $X \rightarrow Y(c\%)$로 표현한다. 즉 한 트랜잭션이 X를 포함하고 있다면 그 트랜잭션이 Y를 c%의 확률로 포함하고 있다는 의미이다[3].

X의 지지도 support(X)는 전체 트랜잭션에서 X를 포함하는 트랜잭션의 비율로 규칙의 통계적 중요성을 의미하며 규칙 $X \rightarrow Y$의 신뢰도 $conf(X \rightarrow Y)$는 X를 포함하는 트랜잭션에서 Y를 동시에 포함하는 트랜잭션의 비율로 규칙의 강도를 의미하는 척도로 연관 규칙 탐사에 이용된다[3,4,5].

2.2 Apriori 연관 규칙 탐사 기법

연관 규칙 탐사 방법 중의 하나인 Apriori는 지지도를 이용하여 동시에 자주 나타나는 항목(빈발 항목 집합)들을 정제하고 빈발 항목 집합에서 생성된 규칙들은 신뢰도를 이용하여 정제하는 방식이다.

Apriori는 후보 항목 집합에서 각각의 지지도를 계산한 후 사용자가 정의한 지지도보다 크거나 같은 조건을 만

족하는 데이터로 빈발 항목 집합을 구성한다. 그리고 후보 항목 집합은 전 단계의 빈발 항목 집합의 조인연산을 통해 구성된다[4]. Apriori 알고리즘은 AprioriTid, AprioriHybrid 등과 같이 확장되어 연구되고 있다[3,4,6].

2.3 다단계 연관 규칙 탐사 기법

다단계 연관 규칙 탐사는 기존의 연관 규칙 탐사 방법을 확장함으로써 이루어진다[6]. Apriori는 계층이 없는 데이터에 대해 연관 규칙을 탐사하는 방식이다. 따라서 실제 데이터에 계층이 존재한다면 기존의 Apriori 탐사 방법은 수정이 필요하다.

데이터의 계층 분류는 미리 데이터베이스에 분류가 되어 있어야 한다. 데이터의 계층 분류의 예는 그림1과 같다.

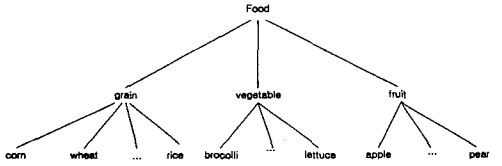

그림 1 .Food의 계층도

다단계 연관 규칙 탐사 방법도 Apriori와 같이 지지도와 신뢰도라는 척도를 사용하여 데이터를 정제한다. 그러나 상위 단계에서 지지도를 만족하지 못하는 데이터들은 하위 단계에서도 지지도를 만족할 수 없으므로 이러한 경우 후보 항목 집합 구성을 하지 않는다.

다단계 연관 규칙에서 하위 단계의 데이터로 구성된 연관 규칙은 상위 단계에서 생성된 규칙보다 명확하고 구체적인 특성을 갖는다. 하위 단계에서 발생하는 연관 규칙을 탐사하기 위해 단계마다 다른 지지도를 적용하여 규칙을 탐사한다. 즉 하위 단계에는 지지도 값을 작게 설정하고 상위 단계에서는 높은 지지도를 적용하는 방법으로 빈발 항목 집합을 탐사한다.

3 빈도수를 고려한 다단계 연관 규칙 탐사 기법

[6]에서는 다단계 연관 규칙 탐사 기법으로 각 단계별로 다른 지지도를 적용하여 하위 단계의 규칙을 탐사하는 방식을 제안하였다.

그러나, 단계별로 지지도를 적용하는 방법은 데이터의 빈도수를 고려하지 못한다는 문제가 있다. 따라서 하위 단계의 데이터나 같은 단계에 있지만 상대적으로 다른 하위 단계 아이템보다 적게 발생하는 데이터와 같이 빈도수가 적은 항목들은 계층별로 적용되는 지지도만으로는 의미 있는 규칙을 발견할 수 없다. 예를 들어 대규모의 할인 매장의 경우라면, 식료품처럼 상대적으로 많이 판매되는 제품과 가전 제품처럼 판매횟수는 식료품보다 적지만 판매 횟수에 비해 많은 이익을 가져다 줄 수 있는 제품이 존재한다. 하위 단계의 아이템들은 상위 단계 아이템보다 발생 빈도수가 적은 것은 분명하지만 앞의 예처럼 식료품의 하위 단계 아이템들은 가전제품의 상위 단계 아이템보다 더 빈발하게 판매될 수 있다. 즉 각 단계별로 지지도를 적용하는 것만으로는 데이터 아이템들 간에 존재하는 의미있는 규칙을 모두 발견할 수 없다.

본 논문에서는 데이터의 빈도수를 고려하여 다단계 연관 규칙 탐사 시 하위 단계의 의미 있는 규칙과 상대적으로 희소한 데이터의 의미 있는 규칙도 찾을 수 있는 MRSApriori(Multiple-level Relative Support Apriori) 탐사 방법을 제안한다.

3.1 상대 지지도

본 논문에서는 데이터 간의 발생에 대한 상대적인 비율인 상대 지지도 Rsup(Relative Support)를 이용하여 다단계 연관 규칙에서의 의미 있는 하위 단계의 규칙과 상대적으로 희소한 데이터에 대한 규칙을 찾을 수 있는 MRSApriori 방법을 제안한다. 상대 지지도는 데이터베이스에서 연관되는 항목들이 데이터의 발생 빈도 중에서 얼마만큼의 비중을 차지하는지 나타내는 척도로써 다음과 같이 정의된다.

【정의 1】 상대 지지도Rsup(Relative Support)
데이터 i_k의 지지도가 sup(i_k)일 때, 데이터 항목 $i_1,i_2, \ldots, i_k$에서의 상대 지지도 Rsup는 데이터의 발생 횟수에 대한 연관된 데이터들의 비율을 표현한 것으로

$$Rsup(i_1,i_2, \ldots, i_k) = \max(sup(i_1,i_2, \ldots, i_k)/sup(i_1), sup(i_1,i_2, \ldots, i_k)/sup(i_2), \ldots, sup(i_1,i_2, \ldots, i_k)/sup(i_k))$$

이다.□

사용자는 연관 규칙 탐사를 위해 지지도와 신뢰도처럼 상대 지지도의 임계값인 최소 상대 지지도(minRsup)를 입력한다. Apriori와 같은 기존의 연관 규칙 탐사 방법은 지지도로 데이터를 정제해나가고 신뢰도로 규칙을 검증하지만 제안하는 MSRApriori 방법은 지지도와 상대 지지도로 데이터를 정제한다. 제안하는 MSRApriori 방법에서는 최소 상대 지지도 값이 0인 경우에 기존의 Apriori와 동일하게 동작하며 최소 상대 지지도의 값이 높을수록 발생 빈도수에 비해 높은 비율로 동시에 나타나는 항목들을 탐사한다.

3.2 MSRApriori의 후보 항목 집합의 구성

MSRApriori의 후보 항목 집합은 minsup(사용자가 정의한 최소 지지도)를 만족하는 항목들과 그렇지 못한 항목들로 나누어 구성한다. minsup를 만족하는 집합을 빈발 항목 집합이라 하고 minsup을 만족하지 못하지만 항목의 Rsup가 minRsup를 만족하는 항목 집합을 준빈발 항목 집합이라고 한다. 준빈발 항목 집합의 후보 항목 집합 구성은 그림 2와 같다.

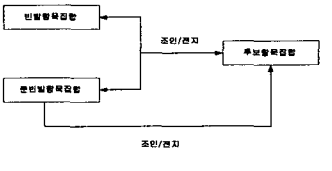

그림 2 . 준빈발 항목 집합에서의
후보 항목 집합 생성 과정

3.3 MSRApriori 알고리즘

다단계 연관 규칙 탐사에서 상대 지지도는 그림3과 같은 방법으로 적용된다. MSRApriori에서는 상대 지지도를 적용함으로써 지지도를 만족하지 않더라도 발생 빈도 중 높은 비율을 가지고 동시에 나타나는 특정 아이템들을 발견할 수 있다.

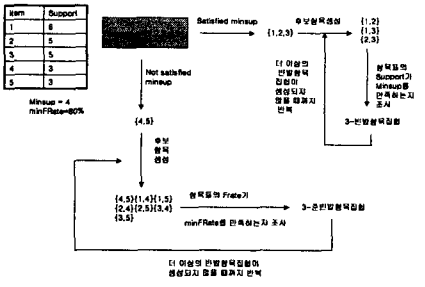

그림 3 .MSRApriori의 과정

제안하는 MSRApriori 방법은 상대 지지도를 이용해 하위 단계의 데이터로 구성된 규칙을 탐사한다. MSRApriori는 상대 지지도를 적용하므로써 비록 발생 빈도는 적지만 높은 비율을 가지고 동시에 나타나는 데이터 사이의 연관 규칙을 탐사할 수 있다. 다단계 연관 규칙 탐사 시 상위 단계, 하위 단계 아이템의 빈도수를 동일하게 보는 기존 방법 대신에 데이터들간의 상대적인 발생 비율을 고려하여 하위 단계를 구성하는 데이터가 이루는 의미 있는 규칙을 탐사할 수 있다. 또한 Apriori와 같은 기존의 연관 규칙 탐사 방법은 같은 단계에 있으나 상대적으로 희소한 성질을 갖는 데이터들은 각각 단계마다 지지도를 적용해도 탐사되기 어렵다. MSRApriori의 상대 지지도는 이러한 경우에도 데이터의 빈도수를 고려하기 때문에 의미 있는 데이터들을 탐사할 수 있다. 본 논문에서 제안하는 MSRApriori의 자료구조는 표2와 같으며 MSRApriori 알고리즘은 <알고리즘1>과 같다.

level	데이터의 단계
minsup	사용자가 정의한 최소 지지도
minRsup	사용자가 정의한 최소 상대 지지도
k-itemset	k개의 아이템으로 구성된 집합
L[level, k]	level에서의 k-빈발항목집합
NL[level,k]	level에서의 minRsup를 만족하는 k-준빈발 항목 집합
C_k	k- 후보항목집합
NC_k	minRsup를 만족하는 원소들에 k- 후보항목집합
apriori-gen	후보항목생성 함수
subset	트랜잭션에 후보항목이 존재하는지 검사하는 함수
LL[level,k]	모든 단계별로 수집된 최종 빈발 항목 집합

표 2 . 알고리즘에 사용되는 표현들

■ input : minsup, minRsup
■ output : large itemset

L_1 = { satisfied minsup 1-itemset}
NL_1 = {L_1을 제외한 item}

for (level=1; L[level,1]≠Ø or level<max_level;level++) do
  for (k=2; L[level, k-1] ≠ Øor NL[level,k-1] ≠Ø;k++)
    C_k = apriori_gen(L[level, k-1]);
    NC_k= apriori_gen(NC_{k-1}) + apriori_gen(NC_k, L[level,k-1])
    for all transaction t∈ D do
      C_t = subset(C_k ,t);
      for all candidates c ∈ C_t do
        c.count++
      end
      if c.count ≥ minsup then
        L[level,k] ={c∈C_t | c.count ≥ minsup}
      else
        item i_k ∈C_t do
          c.Rsup =
              max(sup(i_1,i_2,...,i_k)/sup(i_1),
                  sup(i_1,i_2, ... , i_k)/sup(i_2),
                  ...,
                  sup(i_1,i_2, ... , i_k)/sup(i_k))
          if c.Rsup ≥ minRsup then
            NL[level,k] = {c ∈ C_t | c.Rsup ≥ minRsup}
        end
      end
    end
  LL$_k$ = L[level,k] + NL[level,k]
end
Answer = $\cup_k$LL[level, k];

알고리즘 1. MSRApriori 알고리즘

4 결 론

본 논문에서는 다단계 연관 규칙 탐사 시 상대 지지도라는 척도를 추가 적용하여 하위 단계와 상위 단계간의 다양한 데이터간의 상대적인 빈도수를 고려한 규칙 탐사를 할 수 있는 MSRApriori 알고리즘을 제안하였다. MSRApriori 알고리즘은 상대 지지도를 통해 다단계 연관 규칙 탐사에서도 지지도를 만족하지 못하더라도 데이터의 발생 빈도수에 비해 특정 아이템들과 상대적으로 높은 비율로 동시에 나타나는 아이템들을 탐사할 수 있다.

향후 연구 방향으로 후보 항목 집합의 축소, 시스템 구현, 효율적 알고리즘 개선 등의 연구가 필요하다.

5 참고문헌

[1]. Jiawey Han, "Data Mining", J.Urban and P.Pasgupta. Encyclopedia of Distributed Computing, Kluwer Academic Publisher

[2]. Heikki Mannila, "Methods and problems in data mining", Proceedings of International Conference on Database Theory , 1997

[3]. 박종수,유원경,홍기형, "연관 규칙 탐사와 그 응용", 정보과학회지,98년 9월

[4].Rakesh Agrawal, Ramakrishnan Srikant. "Fast Algortihms for Mining Association Rules", Proceedings of VLDB conference, 1994

[5].Rakesh Agrawal, Tomasz Imielinski, Arun Swami, "Mining Association Rules between Sets of Items in Large Database", Proceeding of ACM SIGMOD, 1993

[6]. Jiawey Han, Y.Fu, "Multiple-level association rules from large databases", Proceeding of VLDB, 1995.9