2001년 한국정보처리학회 추계 학술발표논문집 제8권 제2호

항목의 개체수를 이용한 확장된 데이터 마이닝 연관규칙

조형진, 황병연
가톨릭대학교 컴퓨터공학과
e-mail:hjcho7@hanmail.net, byhwang@www.cuk.ac.kr

Extended Association Rules of Data Mining using Number of Items

Hyoung-Jin Cho, Byung-Yeon Hwang
Dept. of Computer Engineering, The Catholic University of Korea

요약

현 시대에 살아가는 사람들은 정보의 홍수 속에서 살아간다고 해도 과언이 아니다. 컴퓨터 시스템의 발달과 데이터베이스 시스템의 사용의 증가로 컴퓨터에 저장되는 정보의 양은 폭발적으로 증가하고 있다. 현재의 컴퓨터에 저장되어 있는 대용량 데이터베이스에는 사용자가 미처 파악하지 못하는 중요한 정보가 포함되어 있을 수 있다. 본 논문에서는 데이터 상호간의 연관규칙에서 각 항목의 개체수를 고려하여 사용자들에게 좀 더 유용하고 다양한 종류의 데이터를 제공하기 위해 새로운 데이터 마이닝 연관규칙 방법을 제안한다.

1. 서론

최근 컴퓨터 시스템의 발달과 그로 인해 발생되는 많은 양의 데이터로부터 파악되지 않은 정보를 유추해 내는 문제가 대두되고 있다. 정보를 유추하기 위해서는 데이터의 연관성을 고려하는 한 가지 방법이 있다. 그러나 실질적으로 데이터의 연관성을 찾기 위해서는 복잡한 계산이 요구된다. 이러한 문제점은 데이터베이스 스키마를 작성할 때 이미 알려진 자료를 바탕으로 스키마를 설계했기 때문에 발생한다. 기존의 데이터베이스 시스템은 SQL과 같은 일정한 형식의 질의를 벗어난 정보를 검색하는데는 한계를 나타낸다. 이러한 문제점을 해결하기 위해서 데이터베이스 분야와 인공지능의 지식발견 분야의 결합적인 접근 방식으로 대용량 데이터베이스에서 쉽게 발견할 수 없고 알려지지 않은 일정한 패턴이나 정보를 찾아내려는 시도가 90년대 초반부터 시작되어 왔다. 데이터 마이닝(Data Mining)은 대용량 데이터에서 숨겨진 유용한 정보나 패턴을 알아내는 일종의 방법론이라고 할 수 있다. 이러한 연구 중 많은 데이터 속에서 효율적으로 정보를 얻어내기 위해 데이터에 대해 여러가지 작업유형이 연구되기 시작했다. 그 중 하나가 연관규칙이다. 즉, {펜}→{잉크}라는 경우를 예로 들어 규칙을 적용할 때, 이 규칙은 어떤 거래에 있어서 '펜 한 자루를 사게되면 그 거래에 있어서 잉크도 사게될 가능성이 많다.' 라고 해석하는 것이다.

각각의 데이터의 연관성을 고려해서 데이터베이스를 만들면 시간적으로 과거나 현재뿐만 아니라 미래의 상황도 예측 가능하게되어 의사결정시스템에 중요한 역할을 할 수 있다[1][2]. 그러나 지금까지의 연구에서는 데이터 상호간의 관계만을 고려했을 뿐, 데이터 항목의 개체수를 고려하지 않았다. 본 논문에서는 미래의 상황을 예측할 때 각 항목의 개체수를 적용해서 좀 더 유용하고 다양한 종류의 데이터를 얻기 위한 방법을 제안한다.

2장에서는 일반적인 데이터 마이닝 개념과 관련사항을 언급했으며 3장에서는 확장된 데이터 마이닝 연관규칙을 제안하였다. 4장에서는 본 논문에서 제안한 ExAR의 전개 사항을 보이며, 5장에서는 기존의 연관규칙의 결과에 ExAR을 적용하여 결과를 비

교하고, 6장에서 결론을 맺는다.

2. 관련연구

데이터 마이닝이란 자동화되고 지능을 가진 데이터베이스 분석기법으로 90년대 초반부터 지식발견(KDD: Knowledge Discovery in Database), 정보발견(Information Discovery), 정보수확(Information Harvesting)등의 이름으로도 소개되어 왔다. 데이터 마이닝은 일반적으로 대량의 데이터로부터 새롭고 의미있는 정보를 추출하여 의사결정에 활용하는 작업이라 정의된다. 정보를 추출하는 과정에서 연관규칙은 데이터간의 상호관계를 고려해야 한다. 즉, 장래의 행위에 지침을 얻고자 큰 데이터 집합에 들어있는 흥미로운 경향이나 패턴을 찾아내서 최소한의 사용자 입력으로 데이터 내에 숨어 있는 흥미로운 정보들을 얻어내는 것이다.

실제 세계에서 데이터 마이닝이란 여러 가지 알고리즘 중 하나를 단순히 적용하는 것 뿐만 아니라 무가치한 데이터를 교정하고 그 활동을 통해 패턴들을 찾아내고 각 패턴들의 신뢰도를 높이는 것이다.

2.1. 데이터 마이닝 조회방식

데이터 마이닝은 가설을 발견하는 방식으로 사용자의 부가적인 입력을 거의 받지 않고 정보를 찾는다. 이러한 데이터 마이닝 설계는 데이터로부터 짧은 시간 내에 가능한한 다수의 유용한 가설을 산출해내는 방식으로 정보를 발견하도록 설계되어 있다. 그림 1은 데이터 마이닝을 적용한 지식발견 체계도이다[3].

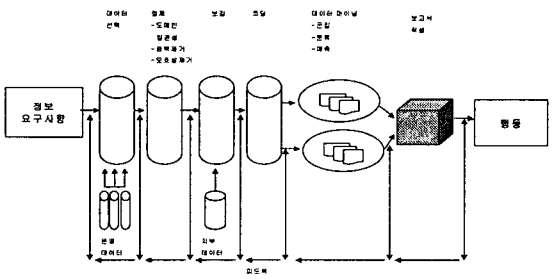

그림 1 데이터 마이닝 조회방식

2.2. 연관규칙

X와 Y를 항목들의 집합이라고 하면 연관규칙은 R: X->Y형식을 갖는다. 이때 X와 Y는 서로 다른 항목 집합이다. 만일 한 트랜잭션이 X를 지지한다면, 확률적 근거에 의해 Y도 지지할 것이라는 예측으로 이해될 수 있는 것이 연관규칙이다. 이런 확률을 신뢰도(Confidence)라 한다. R의 신뢰도는 X를 지지하는 트랜잭션(Transaction)에 대하여 Y 또한 지지할 조건부 확률로 정의된다. 즉, conf(R) = supp(X∪Y)/supp(X)이다. supp(X)는 X의 지지도(Support)를 나타내는데, 품목집합 X∪Y에 대한 지지도는 전체 트랜잭션에 대한 X와 Y를 만족하는 트랜잭션의 비율이다. 즉 트랜잭션에서 발생하는 항목의 수를 백분율로 표시한 것을 말한다. 규칙의 신뢰도는 조건부에 대해 결과부가 얼마나 자주 적용될 수 있는지를 나타내고 반면 지지도는 그 규칙 전부가 얼마나 믿을 만한지를 보여준다. 규칙이 데이터베이스에서 적절해지려면 충분한 지지도와 신뢰도를 가져야한다. 그러므로 어떤 주어진 최소 신뢰도 min_conf와 최소 지지도 min_supp에 대해 만일 conf(R)≥min_conf이고 supp(R)≥min_supp라면 규칙은 성립된다[2].

3. 확장된 데이터 마이닝의 연관규칙

기존의 데이터 마이닝의 연관규칙에서는 각 항목의 수량을 고려하지 않고, 각 항목의 트랜잭션의 횟수만을 고려해서 연관성을 결정하였다. 본 논문에서는 항목의 개체수를 고려하여 기존의 데이터 마이닝의 연관규칙을 확장한 ExAR(EXtended Association Rules)를 제안한다. 또한 확장된 연관규칙을 설명하기 위해 본 논문에서 관계계수(Relation Coefficient)와 관계비율(Relation Rate)을 정의한다.

3.1. 관계계수

관계계수는 데이터간의 연관성이 어느 정도인지를 나타낸다. 1-항목일 경우에는 서로 관계된 항목이 없으므로 관계계수를 구하지 않고 2-항목이상의 항목에서는 (식 1)을 이용하여 계산하게 된다.

$$\text{관계계수} = \sum_{j=1}^{l} T_j$$

$$T_j = \left[\frac{\sum_{i=1}^{k} n_i}{MAX(n_1, n_2 \cdots, n_k)}\right] \qquad \text{(식 1)}$$

n_i : 항목의 개수
k : 항목 집합의 항목수
l : 데이터베이스에서 해당 항목 집합이 포함된 트랜잭션의 수

위의 식은 항목이 발생된 트랜잭션에서 각 항목의 개체수가 해당 트랜잭션에서 발생하는 비율을 보이는 식이다.

3.2. 관계비율

관계비율은 전체 데이터베이스에서 데이터 상호간의 지지도에 대해 데이터간의 관계가 어느 정도인지를 나타내는 지표가 된다. 제품간 관계비율이 높다는 것은 제품의 상호 의존도가 높아서 구매의 상승효과를 기대할 수 있는 제품군이라는 것을 의미한다. 관계비율은 다음과 같이 정의된다.

$$\text{관계비율} = \begin{cases} \dfrac{\sum_{i=1}^{m} c_i}{m} & (\text{if } k=1) \quad (\text{식 } 2-1) \\ \dfrac{\text{관계계수}}{\sum_{i=1}^{m} b_i} & (\text{if } k\geq 2) \quad (\text{식 } 2-2) \end{cases}$$

k : 해당 항목 수
m : 트랜잭션의 수
c_i : 트랜잭션에서 발생하는 단일 항목의 개수
b_i : 항목의 개수

4. ExAR의 전개 예

위에서 제안한 관계성을 이용하여 각 단계별로 항목간의 유사정도를 구하여 기존의 방법과 비교한다.

<표 1> 데이터베이스 예

트랜잭션	(항목, 개수)
1	(빵,2), (우유,3), (과자,1), (아이스크림,1)
2	(빵,1), (과자,2), (음료수,1), (아이스크림,2), (우유,1)
3	(우유,1), (과자,1), (껌,1), (사과,1)
4	(빵,2), (우유,1), (과자,1), (음료수,1)
5	(빵,1), (음료수,1), (아이스크림,1), (과자,2)
6	(빵,3), (우유,3), (음료수,1), (과자,1) (아이스크림,1)

<표 1>의 데이터베이스로부터 각각의 항목별 관계비율을 구하여 보면 <표 2>와 <표 3>과 같다.

<표 2> 1-항목 관계비율

후보 1-항목	1-항목 개수	관계비율
빵	9	1.5
우유	9	1.5
과자	8	1.33
음료수	4	0.66
껌	1	0.17
사과	1	0.17
아이스크림	5	0.83

<표 2>의 관계비율은 (식 2-1)을 사용하여 구한 값이다. 예를 들어 빵이라는 항목의 관계비율을 구하여 보면, 빵의 개수(=9)를 데이터베이스의 트랜잭션 수(=6)로 나눈 값(9/6=1.5)이 1-항목 중 빵에 대한 관계비율이다. 다른 항목들도 동일하게 계산한다.

<표 3> 2-항목 관계비율

관계정도 2-항목집합	관계계수	관계비율
빵, 우유	7.17	0.398
우유, 과자	8.16	0.48
과자, 빵	7.33	0.43
빵, 음료수	6.83	0.525
빵, 아이스	6.66	0.476
우유, 음료수	5.33	0.41
우유, 아이스	4.49	0.321
과자, 음료수	7	0.583
과자, 아이스	7	0.538

<표 3>의 관계계수를 구하는 방법으로 (빵, 우유)의 경우를 생각해보면 <표 1>의 데이터베이스로부터 빵과 우유를 구입하는 경우가 네 번이라는 정보를 얻게 된다.

트랜잭션 1- 빵, 빵, 우유, 우유, 우유
트랜잭션 2- 빵, 우유
트랜잭션 4- 빵, 빵, 우유
트랜잭션 6- 빵, 빵, 빵, 우유, 우유, 우유

이 정보로부터 (식 1)을 사용하여 구하게 되면 다음과 같다.

· 트랜잭션 1에서 빵과 우유는 각각 2, 3개씩이다.

$$T_1 = \frac{2+3}{MAX(2,3)} = \frac{5}{3} = 1.67\text{이다.}$$

· 트랜잭션 2에서 빵과 우유는 각각 1, 1개씩이다.

$$T_2 = \frac{1+1}{MAX(1,1)} = \frac{2}{1} = 2\text{이다.}$$

· 트랜잭션 4에서 빵과 우유는 각각 2, 1개씩이다.

$$T_3 = \frac{2+1}{MAX(2,1)} = \frac{3}{2} = 1.5\text{이다.}$$

· 트랜잭션 6에서 빵과 우유는 각각 3개씩이다.

$$T_4 = \frac{3+3}{MAX(3,3)} = \frac{6}{3} = 2\text{이다.}$$

따라서 (식 1)을 사용하여 빵과 우유의 관계계수를 구하면 $\sum_{j=1}^{4} T_j = 1.67+2+1.5+2 = 7.17$이 된다.

다음으로 빵과 우유에 대한 관계비율은 빵과 우유의 항목 개수의 합인 9+9=18로 빵과 우유의 관계계

수를 나눈다. (식 2-2)에 적용하면 7.17/18=0.398이다. 다른 2-항목도 위와 같은 방법으로 구한다.

5. 비교

기존의 연관규칙의 결과에 본 논문에서 제안된 관계비율을 적용하면 다음과 같다.

① 기존의 데이터 마이닝 연관규칙으로 지지도와 신뢰도를 구하면 <표 4>와 같다[4][5].

<표 4> 최소지지도가 0.5이상이고 최소신뢰도가 0.8이상인 항목들

빈발항목	최소지지도(=0.5)	최소신뢰도(=0.8)
1-항목	빵(0.83), 우유(0.83), 과자(0.83), 음료수(0.66), 아이스크림(0.66)	
2-항목	**(빵, 우유)=(0.66)**, (과자, 빵)=(0.83), **(빵, 음료수)=(0.66)**, (우유, 음료수)=(0.5), **(빵, 아이스크림)=(0.66)**, (우유, 아이스크림)=(0.5), (아이스크림, 음료수)=(0.5), (우유, 과자)=(0.83), (과자, 음료수)=(0.66), (아이스크림, 과자)=(0.66)	**(빵→우유)=(0.8)**, (빵→과자)=(0.8), **(빵→음료수)=(0.8)**, **(빵→아이스크림)=(0.8)**, (우유→빵)=(0.8)
3-항목	(빵, 우유, 과자)=(0.66), (빵, 우유, 음료수)=(0.5), (빵, 우유, 아이스크림)=(0.5), (우유, 과자, 아이스크림)=(0.5)	(과자→아이스크림→우유)=(1), (과자→아이스크림→음료수)=(1)

② 데이터의 관계비율을 구하면 <표 5>와 같다.

<표 5> 관계비율

빈발항목	관계비율
1-항목	빵(1.5), 우유(1.5), 과자(1.33), 음료(0.66), 아이스크림(0.83)
2-항목	**(빵, 우유)=(0.398)**, (과자, 음료수)=(0.583), **(빵, 아이스크림)=(0.476)**, (빵, 과자)=(0.43), (우유, 과자)=(0.544), (우유, 음료수)=(0.41), (우유, 아이스크림)=(0.321), (과자, 아이스크림)=(0.538), **(빵, 음료수)=(0.525)**
3-항목	(빵, 우유, 과자)=(0.320), (빵, 우유, 음료수)=(0.349), (빵, 우유, 아이스크림)=(0.289), (우유, 과자, 아이스크림)=(0.280)

<표 4>의 내용을 보면 **(빵, 우유)=(0.66)**, **(빵, 음료수)=(0.66)**, **(빵, 아이스크림)=(0.66)**의 지지도가 동일하고 **(빵→우유)=(0.8)**, **(빵→음료수)=(0.8)**, **(빵→아이스크림)=(0.8)**의 신뢰도가 동일하다. 그렇다면 신뢰도에서 빵이라는 항목을 기준으로 우선순위를 어떻게 정해야하는가 하는 문제가 발생한다. 이에 대한 대안으로 각 항목의 관계정도를 계산한 관계비율을 이용한다. <표 5>의 관계비율을 살펴보면 **(빵, 우유)=(0.398)**, **(빵, 아이스크림)=(0.476)**, **(빵, 음료수)=(0.525)**라는 결과를 볼 수 있다. 따라서 빵을 중심으로 음료수를 진열하는 것이 우유나 아이스크림을 진열하는 것보다 관계비율이 높게 된다. 이 결과는 기존의 방법과는 달리 수량을 고려하여 동일한 조건일 경우 관계비율이 높은 순서로 상품간의 우선순위를 구할 수 있다.

6. 결론

본 논문에서는 기존의 데이터 마이닝 연관규칙에 항목의 개체수를 고려하여 확장시켰으며 그에따라 사용자들에게 보다 정확한 정보를 제공할 수 있는 데이터간의 관계를 파악하는 방법을 제안하였다.

향후 과제로는 관계비율이 동일할 경우 항목 중 어떤 것을 우선해야하는가 하는 문제를 해결해야 하고, 또한 관계비율에 따라 사용자에게 제공되는 정보량에 변화를 주는 새로운 방법등이 연구되어야 한다.

참고문헌

[1] Mohammed J. Zaki, "Parallel and Distributed Association Mining: A Survey," IEEE Concurrency, October-December 1999, pp.14-25.

[2] Rakesh Agrawal and Ramakrishnan Strikant, "Fast Algorithms for Mining Association Rules," IBM Almaden Research Center, 1999.

[3] Pieter Adrianns and Dolf Zantinge, Data Mining, Addison Wesley, 1996.

[4] Albert Y. Zomaya and Tarek EI-Ghazawi, "Parallel and Distributed Computer for Data Mining," IEEE Concurrency, October-December 1999, pp.11-13.

[5] Heikki Mannil, "Methods and Problems in Data Mining," Univ. of Helsinki, 1998.