제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

시정보 반영을 통한 연관규칙의 신뢰도 측정

옥지웅*, 백주련*, 김응모*
*성균관대학교 컴퓨터공학과
e-mail : {okjwguy,wise96,umkim}@ece.skku.ac.kr

Association Rules Reflected Temporal Information

Jeewoong Ok*, Juryon Paik* , Ungmo Kim*
*Dept of Computer Engineering, Sungkyunkwan University

요 약

연관규칙 (Association rule) 마이닝은 무수히 많은 데이터로부터 유용한 정보만을 뽑아내어 실생활에 적용하여 이점을 얻게 하는 데이터마이닝의 가장 핵심적인 연구분야이다. 마켓 기반 데이터들로부터 고객들의 구매유형을 분석하여 적절한 판매전략을 세우거나 기업 데이터로부터 특정 업무와 관련된 의사결정을 지원하는 등의 일이 모두 연관규칙을 기반으로 한다. 그러나 대부분의 연관규칙들은 시간을 고려하지 않는 않거나, 순차패턴만을 고려해왔다. 따라서 하루중 특정 규칙이 발생되지 않는 시간대에도 그 규칙에 대한 불필요한 노력이 있었다. 본 논문에서는 추출된 연관규칙들과 각 트랜잭션에 부여한 시간 정보를 분석하여 특정 항목 (Item) 집합들 간의 연관규칙이 빈번하게 발생하는 시간대를 추출한다. 추출된 시간 정보를 이용하여 시간대별 유용한 판매 전략을 세움으로써, 상품 판매를 극대화하고자 한다.

1. 서론

데이터 마이닝이란 대량의 데[1]이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정으로 여러 기법들이 있다. 이 기법중 연관규칙 마이닝은 현실세계에서 발생하는 대용량의 사건들이 저장되어 있는 데이터베이스내의 항목 집합으로부터 그들 간의 특정 연관성을 찾아내는 작업이다. 이러한 연관규칙 마이닝은 생성된 많은 정보들 중 가치 있는 지식을 얻기 위하여 트랜잭션이 발생한 업무 분야의 특성을 효과적으로 파악할 수 있는 도구로써, 의사결정(decision making)이나 동향 분석, 수요 예측, 시장전략 수립, 상품진열, 의료진단, 바이오 정보학 등의 다양한 분야에서 널리 이용되고 있다. 그러나 기존의 연관규칙 마이닝에 관한 연구는 실세계 데이터를 대상으로 하면서도 시간 개념을 지니지 않은 형태의 데이터 집합을 대상으로 한다. 또한 단순히 순차패턴만을 고려함으로써, 실세계에 유용한 정확한 시간대에 대한 정보를 추출하기에는 어려움이 많았다. 따라서 본 논문에서는 실제 시간데이터를 적용하여 실세계에 유용한 정보를 뽑아내고자 한다.

본 논문에서는 시간 정보가 포함된 확장된 데이터베이스에 대하여 특정 항목들의 집합과 사용자에 의해 주어지는 임계값(최소지지도와 최소신뢰도)을 만족하는 연관규칙을 구하고, 이 연관규칙의 트랜잭션에 따로 부여된 시간 정보(타임스탬프)를 이용하여 시간대별 신뢰도를 구하게 된다. 대상으로 사용되는 데이터는 데이터 마이닝 분야에서 폭넓게 쓰이고 있는 시장바구니 트랜잭션(Market basket transactions)을 이용하며, 각 트랜잭션에 타임스탬프를 추가하여 시간 구간에 대한 신뢰도를 측정한다. 시간대별 신뢰도는 특정 시간대에 연관규칙의 발생빈도를 나타내는 비율로, 이 비율을 이용하여 특정 시간대에 적

본 연구는 한국과학재단
특정기초연구(R01-2004-000-10755-0)지원으로
수행되었음.

합한 판매 전략을 수립함으로써, 해당 물품의 판매를 극대화 시킬 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구를 소개하고, 3장에서는 먼저 연관규칙을 정의하고, 시간 정보를 통한 시간대별 신뢰도를 소개하며, 이 신뢰도가 실세계에 미칠 수 있는 기여도를 분석한다. 마지막으로 4장에서는 결론을 언급한다.

2. 관련 연구

연관규칙은 소매상 데이터베이스에서 소비자의 행동 패턴을 분석하기 위해 Agrawal등에 의해 처음으로 소개되었다.[1] 연관규칙은 대용량 데이터베이스내의 각 항목들의 집합으로부터 그들 간의 특정 연관관계를 암시하는 규칙이다. 연관규칙 마이닝은 데이터베이스에서 추출될 수 있는 모든 항목 집합 중에서 사용자가 정의한 최소 지지도보다 높은 지지도를 지닌 모든 항목 집합을 찾는다.[1] 연관규칙 탐사를 위해 제안된 알고리즘으로 AIS[1], SETM[7], Apriori[3], AprioriTid[3], AprioriHybrid[3], DIC[8], FP-Growth[10]등이 있다.

순차패턴은 동시에 발생될 가능성이 큰 항목 집합을 찾아내는 연관성측정에서 시간이라는 개념이 포함되어 순차적으로 발생될 가능성이 큰 항목집합을 찾아내는 것이다. 순차패턴을 마이닝하기 위한 알고리즘으로는 AprioriAll[9]과 AprioriSome[9]등이 있다.

이러한 연관규칙 마이닝과 순차패턴 마이닝은 시간을 전혀 고려하지 않거나, 시간에 대한 순서만을 고려함으로 인해 실세계의 시간 간격을 지니는 데이터에 대해서는 적용하기 곤란하다. 따라서 시간을 고려하는 마이닝 기법이나 알고리즘에 대한 연구가 필요하다. 본 논문에서는 제시하는 시간대별 신뢰도는 각 연관규칙의 실제 시간 간격을 고려하기 때문에 실세계의 시간 간격을 지닌 데이터에 대해서도 적용이 가능하다.

3. 시간 구간별 신뢰도 측정

3.1 연관규칙

Agrawal 등에 의해 정립된 연관규칙은 X⇒Y 형태로 X는 몸체(*body*), Y는 머리(*head*)로 명명한다. X, Y는 데이터베이스 D를 이루는 트랜잭션들을 구성하는 아이템들의 집합 I의 부분 집합이 된다. 이때 X와 Y 간의 교집합은 존재하지 않는다. 연관규칙은 두 가지의 측정인자를 계산해서 추출된다. 첫 번째 인자는 지지도(*support*)로서, 이 값은 특정 항목 집합의 통계적 중요성을 나타내는 수치로, 데이터베이스(D)에서 X와 Y를 모두 포함하고 있는 트랜잭션들에 대한 비율이다. 두 번째는 신뢰도(*confidence*)이다. 신뢰도는 연관규칙의 강도를 나타내는 척도로서, X가 속하는 트랜잭션들에서 Y 또한 속하는 트랜잭션들에 대한 비율이다. 정형적으로 정의하면 아래와 같다. 함수 freq(X,D)는 집합 D에서 X를 포함하고 있는 트랜잭션들의 비율을 나타낸다고 정의[2]할 때,

$$\text{지지도}(X \Rightarrow Y) = freq(X \cup Y, D)$$

$$\text{신뢰도}(X \Rightarrow Y) = \frac{freq(X \cup Y, D)}{freq(X, D)}$$

유용한 연관규칙 이론은 사용자에 의해 주어진 최소지지도(minimum support : MinSup)와 최소신뢰도(minimum confidence : MinConf)를 임계값으로 한 후 해당 임계값이상이 되는 지지도와 신뢰도를 갖는 특정 규칙들을 지칭한다.[1]

3.2 시간 구간별 신뢰도

시간 구간별 신뢰도는 연관 규칙이 추출되었을 때, 그 연관 규칙에 대해 각 시간별 발생빈도의 비율을 계산한 값이다. 시간 구간별 신뢰도를 측정하기 위해 본 논문에서는 시장바구니 트랜잭션을 이용한다.<표 1> 이를 기반으로 각 트랜잭션에 타임스탬프를 추가한 <표 2>를 대상으로 신뢰도를 측정한다. 여기서 타임스탬프t_{T_i}는 트랜잭션T_i에 할당된 시간이며, 값은 1에서 24까지이다.($1 \le t_{T_i} \le 24$)

<표 1> 시장바구니 트랜잭션의 예

트랜잭션 번호(TID)	항목
1	{빵, 우유}
2	{빵, 기저귀, 맥주, 달걀}
3	{우유, 기저귀, 맥주, 콜라}
4	{빵, 우유, 기저귀, 맥주}
5	{빵, 우유, 기저귀, 콜라}

<표 2> 시간이 들어간 시장바구니 트랜잭션의 예

트랜잭션 번호(T_i)	타임스탬프 (t_{T_i})	항목
1	4	{빵, 우유}

2	5	{빵, 기저귀, 맥주, 달걀}
3	5	{우유, 기저귀, 맥주, 콜라}
4	6	{빵, 우유, 기저귀, 맥주}
5	6	{빵, 우유, 기저귀, 콜라}

[정의 1] 시간에 대한 신뢰도

시간대별 신뢰도(CT_i: Confidence of temporal duration i)는 추출된 유효한 연관규칙들에 부여한 타임스탬프를 이용하여 계산한다. 이 값은, 추출된 연관규칙들이 나타나는 모든 트랜잭션들에서 특정 시간대에만 나타나는 트랜잭션들의 비율이다. 계산식은 다음과 같다.

$$D = \{T_i | i \in [1,n],\ n\text{은 전체 트랜잭션의 수}\}$$

$$D_i = \{T_j | t_{T_j} = i,\ t_{T_j}\text{는 } T_j\text{의 타임스탬프}, i \in [1,24]\}$$

$$CT_i(X \Rightarrow Y) = \frac{freq(X \cup Y, D_i)}{freq(X \cup Y, D)}$$

시간대별 트랜잭션을 나타내는 D_i에서 i는 1에서 24까지의 값을 갖게 되는데, 이는 적용하는 마켓이 24시간 영업한다는 가정이고, 영업시간에 따라 i의 값은 변경이 가능하다.

[예 1] <표 2>에서 연관규칙 {기저귀} → {맥주}를 고려해보자. 최소지지도와 최소신뢰도를 각각 0.5로 설정한다. 전체 트랜잭션에서 {기저귀, 맥주}가 나타난 트랜잭션은 3개이다. 따라서 지지도는 3/5 = 0.6이다. {기저귀}가 나타난 트랜잭션은 4개이므로 신뢰도는 3/4 = 0.75이다. 이 연관규칙의 지지도와 신뢰도가 최소지지도와 최소신뢰도를 만족하므로 이 연관규칙은 유효하다. 이 규칙의 시간대별 신뢰도를 측정해 보면, 시간 5에서 두 번, 시간 6에서는 한 번 발생하므로 다음과 같이 나타난다.

$$CT_5(\text{기저귀} \Rightarrow \text{맥주}) = \frac{freq(\text{기저귀} \cup \text{맥주}, D_5)}{freq(\text{기저귀} \cup \text{맥주}, D)} = 2/3 = 0.67,$$

$$CT_6(\text{기저귀} \Rightarrow \text{맥주}) = \frac{freq(\text{기저귀} \cup \text{맥주}, D_6)}{freq(\text{기저귀} \cup \text{맥주}, D)} = 1/3 = 0.33$$

시간을 고려하지 않고 추출된 연관 규칙에서의 {기저귀}와 {맥주}는 상품진열정도에만 이용될 수 있지만, 위에서 계산된 시간별 신뢰도를 적용하는 경우, 규칙의 발생빈도가 높은 시간 즉, 5시에 집중적인 판매 전략을 세워서 판매량을 늘릴 수 있다. 또한 규칙이 발생되지 않는 시간때에 판매 전략을 세우는 불필요한 작업을 없앨 수 있다.

3.3 시간대별 신뢰도의 기여도

시간대별 신뢰도는 유효한 연관규칙들의 시간대별 발생비율을 나타낸다. 따라서 유효한 연관규칙이 더 효과적으로 작용하는 시간대를 뽑아낼 수 있다. 예를 들면, 3.2절의 [예 1]에서의 연관규칙{기저귀} → {맥주}는 5시가 6시보다 더 빈번하게 발생한다. 이 값은 기업에서 빈번하게 발생하는 시간대에 노력을 기울일 수 있는 척도를 제공해준다. 대상이 되는 데이터 도메인이 시장데이터이므로 마트를 예로 들면 마트에서 각각의 연관규칙이 효과적으로 작용하는 시간대에 가격 할인등의 판매 전략을 통해 판매량을 극대화할 수 있다. 또한 특정 상품이 팔리지 않고 있는 시간대에 그 상품에 들였던 불필요한 인력과 시간들을 줄이고, 그 인력과 시간을 빈번한 연관규칙의 상품들로 옮겨 판매량을 극대화할 수 있다.

4. 결론

연관규칙 마이닝은 대용량 데이터베이스 상에서 단순 통계적 분석으로는 얻기 어려운 새로운 지식을 효율적으로 얻을 수 있도록 도와주는 기법이다. 기존의 마이닝 기법이나 알고리즘들은 시간을 배제하고 관계형 데이터베이스 상에서 연관규칙을 추출하거나, 순차 패턴을 분석하는 기법이 많았다.

본 논문에서는 기존의 관계형 데이터베이스에 시간 정보를 추가한 데이터베이스를 확장하여, 시간에 따른 연관규칙의 빈도를 나타내는 비율로서 시간대별 신뢰도를 제시한다. 이 신뢰도를 통해 시간대별 유용한 판매 전략을 세움으로써, 상품 판매를 극대화시킬 수 있도록 한다.

참고문헌

[1] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases." In Proc. of ACM SICMOD

International Conference on Management of Data, pp.207-216, 1993.

[2] J. R. Paik, H. Y. Youn, U. M. Kim, "A New Method for Mining Association Rules from a Collection of XML Documents" ICCSA 2005, LNCS 3481, pp.936-945, 2005.

[3] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules" In Proc. of the 20th International Conference on Very Large Data Bases, pp.478-499, 1994.

[4] L. Singh, P. Scheuermann, and B. Chen, "Generating association rules from semistructured documents using an extended concept hierarchy" In Proc. of the 6th International Conference on Information and Knowledge Management (CIKM'97), pp.193-200, 1997.

[5] R. Srikant and R. Agrawal, "Mining generalized association rules" In Proc. of the 21st International Conference on Very Large Data Bases, pp.409 -419, 1995.

[6] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables" In Proc. of the 1996 ACM SIGMOD International Conference on Management of Data, pp.1-12, 1996.

[7] M. Houtsma and A. Swami, "Set-oriented mining of association rules" Research Report RJ 9567, IBM Almaden Research Center, Octorber 1993.

[8] S. Brin, R. Motwami, J. Ulman and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data" Proc. SIGMOD 1997.

[9] R. Agrawal, R. Srikant, "Mining Sequential Pattern" Proc. of Int'l Conference on ICDE, March 1995.

[10] J. Han, H. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation" In: Proc. Conf. on the Management of Data (SIGMOD'00, Dallas, TX). ACM Press, New York, NY, USA 2000.