J. of Basic Sciences, Cheju Nat'l Univ.
17(1), 153~158, 2004

기초과학연구 제주대학교
17(1), 153~158, 2004

Apriori 알고리즘에서 연관 규칙 생성

강 형 창 · 김 철 수 · 박 경 린
제주대학교 자연과학대학 전산통계학과

요 약

연관 규칙 마이닝(association rule mining)은 둘 또는 그 이상의 항목들 사이에 존재하는 수많은 연관 규칙들 중에서 지지도(support), 신뢰도(confidence), 향상도(lift)에 근거하여 연관 규칙을 일반화하는 규칙 탐색 방법이다. 연관 규칙은 수많은 항목들 사이에서 후보 빈발 항목을 찾아내어 최소 지지도를 근거로 하여 빈발 항목집합을 찾아낸 후 강한 연관 규칙을 찾아내게 된다. 그러나 빈발 항목집합이 반드시 강한 연관 규칙을 만족하는 것은 아니며, 최소 신뢰도를 만족하는 빈발 항목집합이 강한 연관 규칙을 만족하게 된다. 그러나 최소 신뢰도를 만족하는 빈발 항목집합도 반드시 강한 연관 규칙을 갖고 있다고 할 수 없으며 강한 연관 규칙을 만족하기 위해서는 향상도(lift)를 만족해야 한다. 이에 본 논문에서는 연관 규칙 마이닝 알고리즘중 하나인 Apriori 알고리즘에서 향상도를 이용하여 빈발 항목집합을 생성하고, 연관 규칙을 탐색하는 방법에 대해서 논의한다.

1. 서 론

컴퓨터 시스템의 발달과 데이터베이스 시스템 사용의 증가로 컴퓨터에 저장되는 데이터의 양은 방대해져 가고 있다. 방대한 양의 데이터로부터 유용한 정보를 얻어내기 위한 여러 가지 방법들을 사용하는 기법이 데이터 마이닝(data mining)이며, 그 중 데이터를 탐색하여 규칙을 찾아내는 방법이 연관 규칙 마이닝(association rule mining)이다. 즉, 연관 규칙 마이닝은 둘 또는 그 이상의 항목들(items) 사이에 존재하는 수많은 연관 규칙들 중에서 강한 연관 규칙을 찾기 위해 지지도(support), 신뢰도(support), 향상도(lift)에 근거하여 강한 연관 규칙을 찾아낸 다음 일반화하는 방법이다.

현재까지 연관 규칙 마이닝 알고리즘은 데이터베이스에 저장되어 있는 항목들에서 빈발 항목집합을 찾아낸 후 강한 연관 규칙을 찾아낸다. 이러한 연관 규칙 마이닝 알고리즘들은 다음과 같이 나눌 수 있는데, 하나는 빈발 항목집합을 생성하기 위해 후보 항목집합을 구성한 후 빈발 항목집합을 결정하는 방법이 있으며, Agrawal 등(1994)이 제안한 후보 항목집합들을 구성하고 후보 항목집합들의 발생빈도 수를 계산하여 사용자가 정의한 최소 지지도를 기초로 빈발 항목집합을 결정하는 Apriori 알고리즘을 비롯하여, Liu 등(1999)은 후보 항목집합들을 효율적으로 작게 구하여 이를 기초로 전체 트랜잭션의 크기와 개수를 줄이는 방법인 DHP (Direct Hashing and Pruning)알고리즘 등이 있다. 그리고 다른 하나는 후보 항목집합을 구성하지 않고 빈발 항목집합을 결정하는 방법으로, Park 등(1995)은 데이터베이스를 중복되지 않는 크기로 분할한 후 한번에 한 개의 분할 영역만을 고려하여 그 안에서 빈발 항목집합을 생성하는 Partition 알고리즘을 제안하였으며, Toivonen(1996)는 무작위로 선정된 표본을 이용하여 빈발 항목집합 찾은 후 그 결과를 데이터베이스의 나머지 부분에 적용하여 증명하는 방법인 Sampling 알고리즘을 제안하였다. 그리고 Cheung 등(1996)은 갱신된 데이터베이스에서는 이전에 빈발 항목을 다루었던 항목집합은 데이터베이스 스캔을 생략하는 FUP 알고리즘 등이 있다.

앞서 보듯이 연관 규칙 마이닝은 빈발 항목집합을 찾은 후 강한 연관 규칙을 찾게 된다. 이에 본 논문에서는 후보 항목집합을 구성한 후 빈발 항목

집합을 찾는 Apriori 알고리즘에 향상도를 사용하여 후보 항목집합을 적게 구성하고, 빈발 항목집합을 찾는 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 Apriori 알고리즘에 대하여 설명하고, 3장에서는 Apriori 알고리즘에 대한 문제를 정의하고 제안된 알고리즘을 설명한다. 4장에서는 예제를 통해 Apriori 알고리즘과 제안된 알고리즘을 비교하고, 마지막 5장에서는 결론 및 연구 과제를 다룬다.

2. 연관 규칙 기본 개념과 Apriori 알고리즘

2.1 연관 규칙 기본 개념

k개로 이루어진 항목들의 집합 $I=\{i_1, i_2, \cdots, i_k\}$이 주어지면, 트랜잭션 T는 I의 부분집합으로 정의된다($T \subseteq I$). 이때, 각 트랜잭션들은 중복된 항목을 허용하지 않으며, TID라 불리는 고유한 트랜잭션 아이디를 갖는다. 만일 트랜잭션 T가 X의 모든 항목들을 포함한다면($X \subseteq T$), T가 항목집합 X를 지지한다(support)고 한다. 연관 규칙을 구성하기 위해서는 모든 항목집합들 중에서 후보 항목집합을 구성한 다음, 후보 항목집합에서 빈발 항목집합을 찾기 위해 지지도를 계산한다. 빈발 항목집합이 구성되면 신뢰도를 바탕으로 연관 규칙을 구성한다. 연관 규칙 $R: X \Rightarrow Y$에서 사용되는 용어들은 다음과 같다.

- 트랜잭션(transaction): 발생된 데이터를 저장하는 단위이며, 여러 가지 항목들을 가질 수 있다.
- 항목집합(itemset): 각 개별 트랜잭션에 포함된 단일 항목 또는 복수 항목의 집합
- 후보 항목집합(candidate itemset): 개별 항목집합의 결합을 통해 생성된 항목집합으로써 후보 항목집합에 대한 지지도 및 신뢰도를 계산하여 최소 지지도 및 최소 신뢰도를 만족하는 경우 빈발 항목집합으로 간주한다.
- 빈발 항목집합(frequent itemset): 후보 항목집합에서 최소 지지도 및 최소 신뢰도를 만족하는 항목집합
- 지지도(support): $Sup(X \Rightarrow Y) = P(X \cup Y)$
- 신뢰도(confidence):

$$Conf(X \Rightarrow Y) = P(Y|X) = \frac{P(X \cap Y)}{P(X)}$$

- 향상도(lift):

$$Lift(X \Rightarrow Y) = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)}$$

여기서, $X \subseteq T$, $Y \subseteq T$, $X \cap Y = \emptyset$이다.

2.2 Apriori 알고리즘

Apriori 알고리즘은 빈발 항목집합 특성인 사전 지식(priori knowledge)을 사용한다. Apriori는 k번째 항목집합이 k+1번째 항목집합을 발견하기 위해 레벨단위로 진행하여 반복 접근한다. 첫째로 빈발 1-항목집합을 찾는다. 이 집합을 L_1으로 나타낸다. L1은 2-항목집합인 L_2를 찾는데 사용되며, 이것은 다시 L_3를 찾는데 이용되는 식으로 계속되어 더 이상의 빈발 k-항목집합이 없을 때까지 진행된다. 각 L_k를 찾기 위해서는 전체 데이터베이스에 대한 스캔이 요구된다.

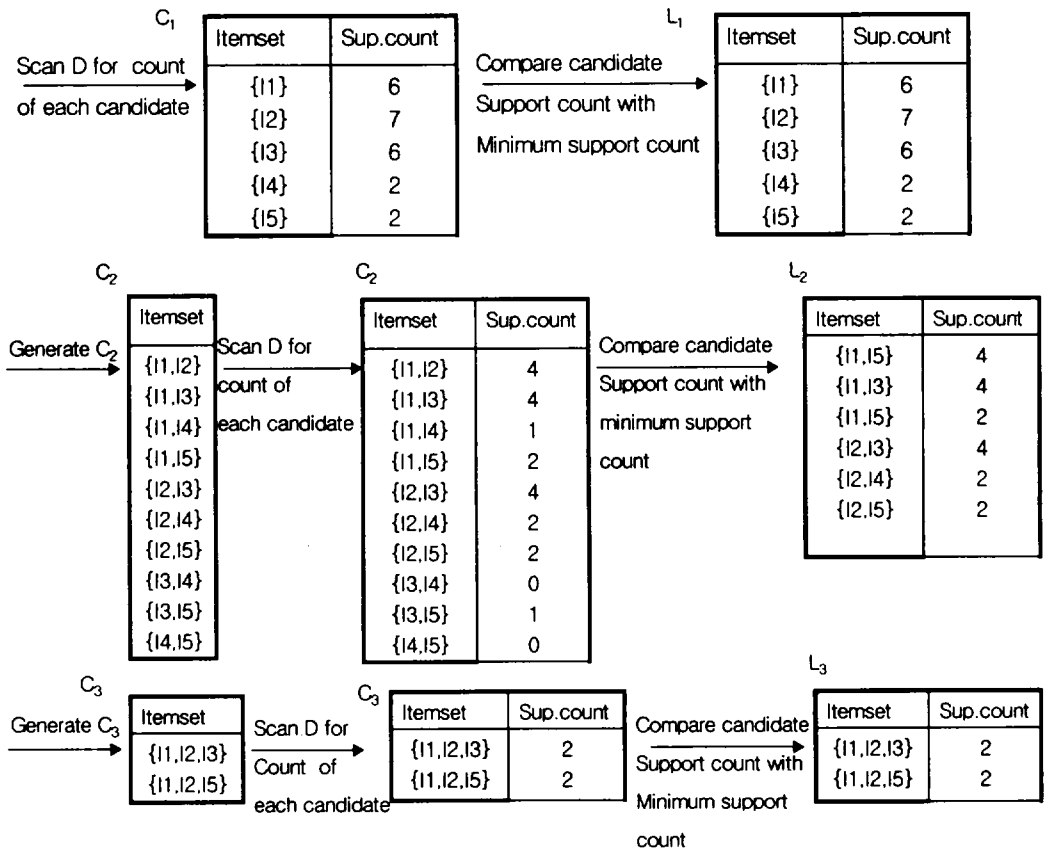
빈발 항목집합을 레벨단위로 생성하는 것을 효과적으로 개선하기 위해서 Apriori 특성인 모든 공집합이 아닌 빈발 항목집합의 부분집합은 반드시 빈번하다는 특성을 사용함으로써 탐색 공간을 감소시키는데 사용할 수 있다. Apriori 알고리즘은 결합(join)과 가지치기(prune)의 두 과정으로 이루어진다.

〈표 1〉 트랜잭션 데이터

TID	List of item IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

〈그림 1〉은 〈표 1〉에서 빈발 항목집합을 구성하기 위해 최소 지지도를 2로 가정하고, Apriori 알고리즘을 수행한 결과이다.

결과는 빈발 항목집합 $L_3 = \{\{I_1, I_2, I_3\}, \{I_1, I_2, I_5\}\}$을 구성한다.



〈그림 1〉 Apriori 알고리즘 예

빈발 항목집합 $L_3=\{\{I_1,I_2,I_3\},\{I_1,I_2,I_5\}\}$로부터 연관 규칙은 다음과 같다.

i) $L_3=\{I_1,I_2,I_5\}$

R: $I_1 \wedge I_2 \Rightarrow I_5$, $Conf(I_1 \wedge I_2, I_5)=2/4=50\%$.

R: $I_1 \wedge I_5 \Rightarrow I_2$, $Conf(I_1 \wedge I_5, I_2)=2/2=100\%$

R: $I_2 \wedge I_5 \Rightarrow I_1$, $Conf(I_2 \wedge I_5, I_1)=2/2=100\%$.

R: $I_1 \Rightarrow I_2 \wedge I_5$, $Conf(I_1, I_2 \wedge I_5)=2/6=33\%$

R: $I_2 \Rightarrow I_1 \wedge I_5$, $Conf(I_2, I_1 \wedge I_5)=2/7=29\%$.

R: $I_5 \Rightarrow I_1 \wedge I_2$, $Conf(I_5, I_1 \wedge I_2)=2/2=100\%$

여기서, 최소 신뢰도를 만족하는 연관 규칙은 강한 연관 규칙을 갖는다고 한다.

ii) $L_3=\{I_1,I_2,I_3\}$

향상도가 1보다 작게 나타나, 연관 규칙이 나타나지 않는다.

3. Apriori의 문제점과 제안된 방법

3.1 Apriori의 문제점

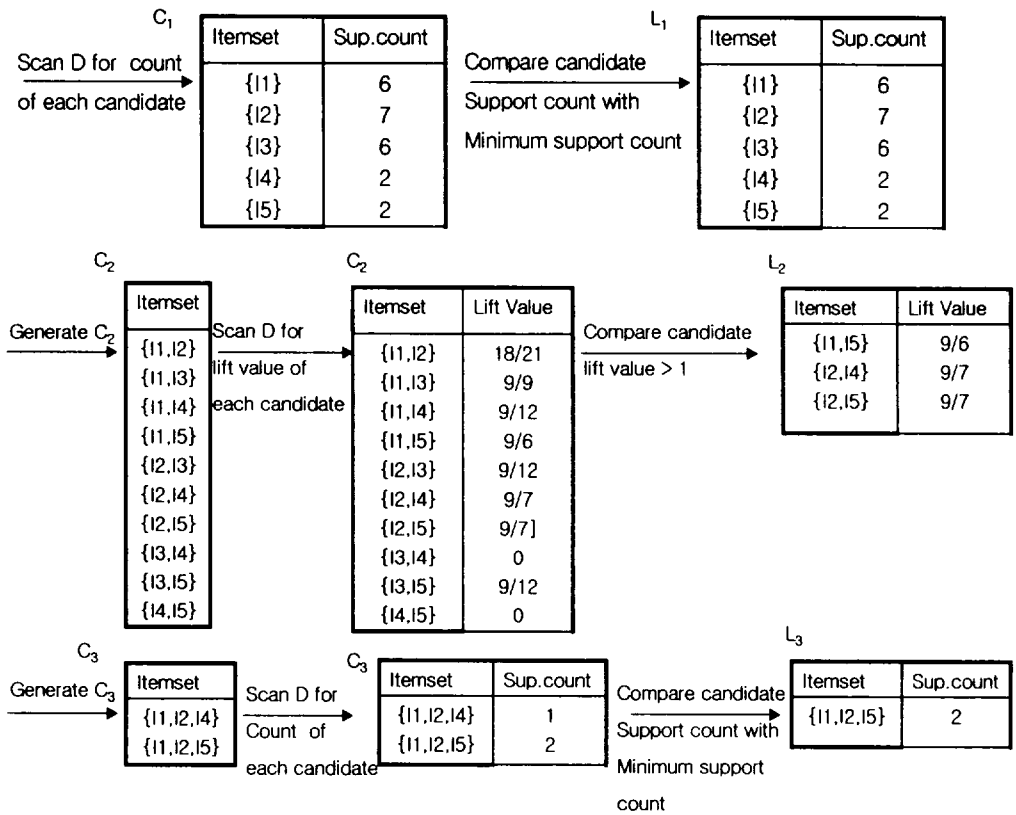
대부분의 연관 규칙 마이닝 알고리즘들은 빈발 항목집합을 생성하는데 특히, Apriori 알고리즘은 빈발 항목집합을 구성하기 위해 후보 항목집합을 먼저 생성한다. 후보 항목집합에서 최소 지지도를 만족하는 항목집합을 빈발 항목집합으로 구성한다. 즉 빈발 항목집합을 구성할 때 가장 큰 영향을 미치는 값은 최소 지지도이며, Apriori 알고리즘은 후보 항목집합에서 최소 지지도를 만족하는지를 확인하기 위해 매번 데이터베이스를 스캔해야 한다. 또한 최소 지지도 값에 따라 빈발 항목집합의 수는 크게 영향을 받는다. 최소 지지도가 너무 크면 빈발 항목집합의 수는 줄어들게 되고, 최소 지지도가

너무 작으면 빈발 항목집합의 수는 커지기 때문에 최소 지지도에 따라 상당한 크기의 후보 집합 생성이 필요할 수 있다. 그리고 빈발 항목집합이 구성된다 하더라도 반드시 강한 연관 규칙이 생성되는 것은 아니다. 그 예로 〈그림 1〉의 결과와 같이 빈발 항목집합 $L_3=\{\{I_1,I_2,I_3\},\{I_1,I_2,I_5\}\}$이 구성되지만, 실제로 강한 연관 규칙은 $L_3=\{I_1,I_2,I_5\}$만 만족한다. $L_3=\{I_1,I_2,I_3\}$는 빈발 항목집합이지만 향상도가 1보다 작게 나타나기 때문에 강한 연관 규칙이 구성되지 않는다.

빈발 항목집합을 구성하는데 가장 큰 영향을 주는 것은 최소 지지도이며, 연관 규칙에 영향을 주는 것은 최소 신뢰도가 된다. 향상도는 지지도와 신뢰도에 의해 결정이 되기 때문에 최소 지지도를 만족하여 빈발 항목집합이 된다하더라도 반드시 연관 규칙이 존재하지 않는 이유는 결국 향상도에 의해 영향을 받기 때문이다.

3.2 제안하는 방법

Apriori 알고리즘에서 빈발 항목집합을 생성하기 위해서 필요한 것은 최소 지지도이다. 그런데 이러한 지지도를 얼마로 할 것인가? 하는 것은 경험적으로 이루어져 왔다. 앞서 설명되었듯이 지지도에 따라 생성되는 빈발 항목집합의 수는 달라지게 되고, 후보 항목집합의 수도 달라지게 된다. 그리고, 생성된 빈발항목 후보 집합에서 빈발항목을 찾아낸 후 최소 신뢰도를 이용하여 연관 규칙을 생성한다. 이에 본 논문에서는 향상도를 도입하여 2번째 빈발 항목집합의 수를 줄임으로써 3번째 후보 항목집합의 수를 줄이는 방법을 제안한다. Apriori 알고리즘에서 빈발 항목집합을 만족하더라도 강한 연관 규칙이 구성되지 않을 수 있으므로 Apriori 알고리즘에 향상도를 도입하여 후보 항목집합 구성을 줄임으로써 빈발 항목집합 생성을 보다 빠르게 하고자 한다. 즉 첫 번째 빈발 항목집합 L_1은 후보 항목



〈그림 2〉 제안된 방법을 이용한 Aprori 알고리즘

집합에서 최소 지지도를 이용하여 구성하고, 구성된 L_1을 이용하여 후보 항목집합을 구성한 후 빈발 항목집합 L_2을 구성하기 위해 각 후보 항목집합에서 최소 지지도가 아닌 향상도를 계산함으로써 빈발 항목집합 L_2의 수를 줄일 수 있다.

4. 적용 예

제안한 방법과 Apriori 알고리즘을 비교하기 위해 〈표 1〉에서 빈발 항목집합을 구성하여 보면 〈그림 2〉와 같다.

각 항목은 후보 항목집합 C_1의 원소이며, 각 항목의 빈도를 계산하기 위해 모든 트랜잭션을 탐색한다. 그 다음 트랜잭션의 최소 지지도를 2라 가정하여 빈발 항목집합 L_1을 결정한다. 빈발 항목집합 L_2를 구성하기 위해 L_1에서 후보 항목집합 C_2를 생성한다. 다음 데이터베이스를 스캔하여 C_2의 향상도를 계산하여 향상도가 1 보다 크면 빈발 항목집합 L_2를 구성한다. 나머지 빈발 항목집합 L_3이후의 항목집합 구성은 Apriori와 같다.

5. 결 론

수많은 연관 규칙에서 강한 연관 규칙을 생성하기 위해서는 빈발 항목집합을 구성한 후에 최소 신뢰도를 이용하여 연관 규칙을 구성한다. 이때 빈발 항목집합을 구성하는 방법은 여러 방법들이 시도되고 있으며 특히 Apriori 알고리즘은 빈발 항목집합을 구성하기 위해 후보 항목집합을 구성하고 최소 지지도를 만족하는 경우에 빈발 항목집합을 구성하게 된다. 이때 최소 지지도에 따라 구성되는 빈발 항목집합의 수는 달라지게 되며, 후보 항목집합의 수도 달라지게 된다. 또한 빈발 항목집합을 만족하더라도 실제로는 강한 연관 규칙을 구성하지 않을 수도 있다. 이에 본 논문에서는 Apriori 알고리즘에서 두 번째 빈발 항목집합을 구성함에 있어 최소 지지도가 아닌 향상도를 도입함으로써 빈발 항목집합의 수가 줄어들고, 후보 항목집합의 수도 줄어듦을 보였다.

참고문헌

[1] An Efficient Algorithm for Updating Discovered Sequential Patterns in Data Mining

[2] Agrawal, R., Srikant R. (1994). Fast algorithms for mining association rules. Proceeding of the 20th VLDB Conference, Santiago, Chile.

[3] Bing, L., Wynne, H., Yiming, M. (1999). Mining Association Rules with Multiple minimum Supports, Proceedings of ACM KDD-99.

[4] Cheung, D.W., Han, J., Ng, V., Fu, A.W., Fu, Y. (1996). A Fast distribution algorithm for mining association rules. Int's Conference on Parallel and Distributes Information System, Miami Beach, Florida.

[5] Jiawei Han, Miceline Kamber, Data Mining: Concepts and Techniques.

[6] Han, J., Kamber M. (2001). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

[7] Park, J.S., Chen, M.S., and Philips, S.Y. (1995). An effecvie hash-based algorithms for mining association rules. Proceedings of ACM SIGMOD conference on Management of Data.

[8] Silverstein, C., Bin, S., Motwani, R. (1997). Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. Data Mining and Knowledge Discovery, No.2 P 39-68.

[9] Toivonen, H. (1996). Sampling Large Database for Association Rules. Proceedings of the 22nd VLDB Conference, Mumbai(Bombay), India.

Association rule generation in the Apriori algorithm

Hyung Chang Kang · Chul Soo Kim · Gyung-Leen Park
Department of Computer Science and Statistics Cheju National University

Abstract. Association rule mining finds interesting association rule among a large set of data itemsets. In general, association rule is determined by a minimum support threshold and minimum confidence threshold. But the strong association rules could not be obtained by such measures. In this article, we discuss the lift value which can be a good method to find frequent itemsets in Apriori algorithm