항목 계층 구조에 기반한 빈발 항목 집합 나열 방법

Item Hierarchy based Frequent Itemset Ordering Method

김 준 우, 강 현 경*
동아대학교, 신라대학교*

Kim jun woo, Kang hyun kyung*
Dong-A Univ., Silla Univ.*

요약

연관 규칙 탐사는 이산적인 항목들을 포함하는 트랜잭션 데이터에 존재하는 항목 간 동시 발생 관계를 찾아내는 데 그 목적을 두고 있다. 연관 규칙은 {전항}→{후항}의 형태를 갖고, 전, 후항은 모두 사전에 정의된 지지도 하한을 만족하는 빈발 항목 집합으로 구성된다. 연관 규칙 탐사에서 문제가 되는 것은 일반적으로 탐사되는 빈발 항목 집합의 개수가 많아지면서 규칙의 개수도 많아지고, 이들 사이에 중복성이 존재한다는 점이다. 따라서 단순히 지지도나 신뢰도 순으로 빈발 항목 집합이나 규칙을 나열하기보다는 항목들의 연관성을 고려하는 것이 분석자에게 보다 도움이 될 수 있다. 본 논문에서는 이를 위하여 연관 규칙 탐사와 함께 계층 군집 분석을 실시하여 항목들 간 연관성을 정리하고, 이를 토대로 빈발 항목 집합들을 나열하는 방법을 제안하고자 한다.

I. 서론

연관 규칙 탐사는 가장 널리 쓰이는 데이터마이닝 기법 중의 하나로, {전항}→{후항} 형태의 규칙을 찾아내는 데 그 목적을 둔다[1]. 일반적으로는 빈발 항목 집합을 탐사하여 이들을 통해 연관 규칙을 생성하는 Apriori 알고리즘이 널리 사용되며, 사전에 분석자가 지지도 하한과 신뢰도 하한을 정해두는 것이 필요하다.

연관 규칙 탐사 결과는 판매대의 제품 배치, 미래 사건의 예측 및 협업적 필터링을 이용한 추천 등에 다양하게 등에 다양하게 활용될 수 있으나, 지지도 및 신뢰도 하한 임계치에 따라 대량의 빈발 항목 집합 및 연관 규칙이 만들어질 수 있다는 단점이 있다. 따라서 이들에 대한 효과적인 관리 및 시각화 방법이 필요할 것으로 생각되고, 이는 최근 분석 대상 데이터의 분량이 방대해지면서 강조되고 있는 인간의 시각적인 데이터 관찰을 도울 수 있을 것으로 생각된다[2].

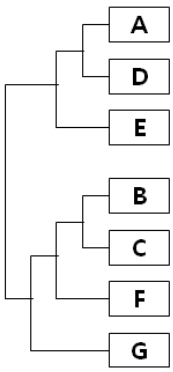
본 논문은 연관 규칙 탐사 과정 중에서 중요한 역할을 하는 빈발 항목 집합들을 나열하기 위하여 종래와 같이 지지도를 기준으로 하기보다 항목 간 연관성을 함께 고려하는 것을 제안한다. 항목 간 연관성을 고려하기 위해서는 병합형 계층 군집 분석을 활용하였다.

Ⅱ. 계층 군집 분석

계층 군집은 본래 데이터의 레코드들 중 유사한 것들을 한데 모으기 위한 군집 분석의 일종으로, 트리 모양의 레코드 계통도가 작성된다는 특징이 있다. 대표적으로는 계통도를 리프에서부터 그려나가는 병합형 계층 군집과 루트에서부터 그려나가는 분할형 계층 군집이 있다. 전통적으로는 계층 군집 역시 여타의 군집 분석 방법들과 마찬가지로 일반 데이터에 포함된 레코드 간 유사도를 유클리드 거리 등으로 계산하여 두 개 레코드를 비교하는 방법을 사용하였다[3].

반면, 트랜잭션 데이터에서는 레코드들을 군집하기보다 데이터를 구성하는 항목들에 대한 군집이 가능한데, 이를 위해서는 항목 집합 A, B 간 유사도를 항목 집합 A∪B의 지지도를 이용하여 측정하는 방법을 사용할 수 있다[4][5].

본 논문에서는 연관 규칙 탐사와 함께 위와 같은 지지도 기반의 계층 군집을 실시하여, 그림 1과 같이 계통도 형태로 빈발 1-항목집합들의 항목 계층 구조를 먼저 파악하는 것을 제안한다.

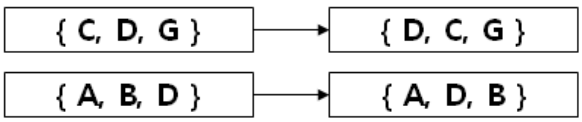


▶▶ 그림 1. 빈발 1-항목 집합 계층 구조

이 과정에서 향후 빈발 항목 집합 나열을 위해 계통도상의 리프로 나타나는 항목들을 먼저 정렬하게 되며, 이러한 정렬은 그림 1과 같은 모양의 계통도의 루트에서 시작하여 매번 트리의 분기점을 만날 때마다 각 분기의 하위에 존재하는 항목들 중 가장 빈발한 것의 지지도를 조사하여, 높은 값을 갖는 분기를 상위에 배치하여 이루어진다. 예를 들어, 그림 1에서는 항목 A, B, E 중 가장 빈발한 것의 지지도가 항목 B, C, F, G 중 가장 빈발한 것의 지지도보다 높은 것을 의미한다.

Ⅲ. 빈발 항목 집합 나열

앞 장에서 설명한 방법으로 빈발 1-항목집합들에 대한 나열 순서가 결정되고 난 후에는 이를 나머지 빈발 항목 집합들을 나열하는데 사용할 수 있다. 이를 위하여 먼저, 각 빈발 항목 집합들이 그림 1과 같은 빈발 1-항목집합들의 순서를 반영할 수 있도록, 그림 2와 같이 빈발 항목 집합에 포함된 개별 항목들의 순서를 재배열한다.



▶▶ 그림 2. 항목집합 재정렬

재배열을 완료한 항목집합들은 이제 빈발 1-항목집합들의 순서에 따라 사전 순으로 나열되며, 나열된 항목집합의 목록은 종래 일반적으로 많은 항목집합을 나열할 때 단순히 각 항목집합의 지지도 또는 사전 순서를 기준으로 삼았던 것과 달리, 개별 항목 간 연관성을 고려하여 서로 연관성이 있는 항목집합이 가까이 배치되도록 하는 것이 가능하다.

Ⅳ. 결론

연관 규칙 탐사는 그 기본적인 개념은 비교적 단순한 편이나, 잦은 데이터 스캔으로 인한 소요 시간 문제, 그리고 탐사 결과에 대한 시각화 등이 중요하게 다루어져야 한다. 본 논문에서는 이를 위하여 빈발 항목 집합들을 나열할 때, 항목 간 관련성을 고려하는 방법을 제안하였으며, 이를 통해 보다 의미있는 형태로 빈발 항목 집합들을 나열하는 것이 가능할 것으로 기대된다.

반면, 그렇다고 하더라도 단순히 빈발 항목 집합들을 텍스트 형태로만 나열한다는 것은 실제 연관 규칙 탐사 결과를 활용하는 데 있어 실질적인 효과가 크지 않기 때문에 향후에는 본 논문에서 제안하는 빈발 항목 집합 나열 방법을 연관 규칙 자체들에 대한 여러 가지 시각화 목적으로 적용하는 연구를 지속할 계획이다.

감사의 글

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (2012R1A1A1044834)

■ 참 고 문 헌 ■

[1] Agrawal, R., Imielinski, T. and Swami, A. "Mining Association Rules between Sets of Items in Large Databases," Proceedings of the 1993 ACM-SIGMOD Conference on Management of Data, pp.207-216, 1993.

[2] de Oliveira, M.C.F. and Levkowitz, H. "From Visual Data Exploration to Visual Data Mining: A Survey," IEEE Transactions on Visualization and Computer Graphics, Vol.9, No.3, pp.378-394, 2003.

[3] Tan, P.-N., Steinbach, M. and Kumar, V. Introduction to Data Mining, Addison Wesley, 2005.

[4] Tsui, C.-J., Wang, P., Fleischmann, K.R., Sayeed, A.B. and Weinberg, A. "Building an IT Taxonomy with Co-occurrence Analysis, Hierarchical Clustering and Multidimensional Scaling," Proceedings of iConference, pp.247-256, 2010.

[5] 김준우, 주지영, 홍성용, 이문용, 윤완철, "효과적인 학습 전략을 위한 컨텐츠 주제어 군집 방법", 한국정보과학회 제37회 추계학술발표논문집, 제37권, 제2호, pp.317-322, 2010.