효율적인 밀집 및 희소 빈발 항목 집합 탐색 방법

이경민, 정석호, 신동문, Ibrahim Musa Ishag Musa, 이동규, 손교용, 류근호
충북대학교 데이터베이스연구실
e-mail: {min9709, sukhojung, mastershin216, ibrahim, dglee, gysohn, khryu}@dblab.chungbuk.ac.kr

An Effective Method for Dense and Sparse Frequent Itemsets Mining

Gyeong Min Yi, Sukho Jung, DongMun Shin, Ibrahim Musa Ishag Musa, Dong Gyu Lee, Gyoyong Sohn, Keun Ho Ryu
Database/Bioinformatics Laboratory
Chungbuk National University

요 약

트리기반 빈발 항목 집합 알고리즘들은 전체적으로 밀집 빈발 항목 집합에는 효율적이고 빠르게 빈발 항목 집합을 탐색하나 희소 빈발 항목 집합에는 효율적이지 않고 빈발 항목 집합을 빠르게 탐색하지 못한다. 반면에 배열기반 빈발 항목 집합 알고리즘은 희소 빈발 항목 집합에 효율적이고 빠르게 빈발 항목 집합을 탐색하나 밀집 빈발 항목 집합에는 효율적이지 않고 빈발 항목 집합을 빠르게 탐색하지 못한다. 밀집 및 희소 빈발 항목 집합 모두 효율적으로 빈발 항목 집합을 탐색 하고자 하는 시도가 있었으나 두 가지 종류의 알고리즘을 동시에 사용하므로 각각의 알고리즘을 사용할 정확한 기준 제시가 어렵고, 두 가지 알고리즘의 단점을 내포한다. 따라서 본 논문에서는 단일 알고리즘을 사용하여 밀집 빈발 항목 집합 및 희소 빈발 항목 집합 모두에 대해 작은 메모리 공간을 사용하면서도 효율적이고 빠르게 빈발 항목 집합을 탐색할 수 있는 CPFP-Tree라는 새로운 자료구조와 탐색 방법을 제안한다.

1. 서론

빈발 항목 집합 탐색은 트랜잭션 데이터베이스(Transaction Database)에서 전체 트랜잭션 개수 대비 항목 집합이 발생한 트랜잭션 개수 비율이 특정 지지도 임계값(Support threshold)보다 큰 지지도를 갖는 모든 항목 집합을 파악하는 작업이다.

빈발 항목 집합 탐색은 연관규칙 마이닝 분야의 가장 핵심적인 부분으로, 많은 연구가 활발히 진행되고 있으며 상호관계(correlations), 인과관계(causality), 순차 패턴(sequential patterns), 에피소드(episodes), 다차원 패턴(multi-dimensional patterns), 최대 패턴(max-patterns), 부분 주기성(partial periodicity), 신흥 패턴(emerging patterns) 등 여러 마이닝 분야에서 다양하게 활용되고 있다[1, 2].

빈발 항목 집합 탐색 방법은 사용하는 자료구조에 따라 배열기반 알고리즘과 트리 기반 알고리즘으로 구분된다. 빈발 항목 집합은 빈발 항목 집합의 특성에 따라 밀집도가 낮은 희소 빈발 항목 집합과 밀집도가 높은 밀집 빈발 항목 집합으로 구분된다[1, 2, 3, 4, 5, 6, 7, 8].

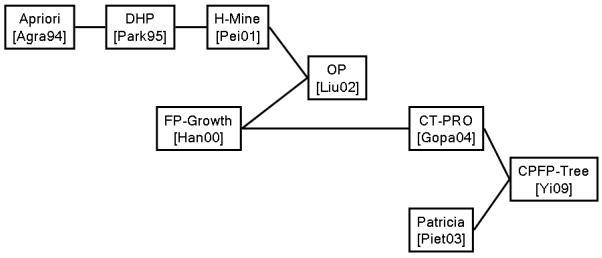

그림 1 빈발 항목 집합 알고리즘 발전도

그림 1은 종래의 빈발 항목 집합 탐색 알고리즘 발전을 나타낸 도면으로서, 자료구조에 따라 배열기반과 트리기반으로 분류하고 발전 순서에 따라 표현한 것이다[2, 3, 4, 5, 6, 7, 8].

트리기반 알고리즘들은 전체적으로 밀집 빈발 항목 집합에는 효율적이고 빠르게 빈발 항목 집합을 탐색하나 희소 빈발 항목 집합에는 효율적이지 않고 빈발 항목 집합을 빠르게 탐색하지 못한다[2, 3, 5]. 반면에, 배열기반 알고리즘은 희소 빈발 항목 집합에 효율적이고 빠르게 빈발 항목 집합을 탐색하나 밀집 빈발 항목 집합에는 효율적이지 않고 빈발 항목 집합을 빠르게 탐색하지 못한다[4, 7, 8].

OP 알고리즘은 능동적으로 밀집도가 높으면 FP-Growth 알고리즘을 사용하고 밀집도가 낮으면 H-Mine을 사용한

이 논문은 2009년 교육과학기술부로부터 지원을 받아 수행된 연구(지역거점연구단육성사업/충북BIT연구중심대학 육성사업단)와 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R11-2008-014-02002-0)

다[6]. 따라서 밀집 및 희소 빈발 항목 집합 모두 효율적으로 빈발 항목 집합을 탐색한다. 그러나 밀집도에 대한 명확한 기준 제시가 어려우며, 밀집도가 높아 FP-Growth 알고리즘을 사용할 경우 FP-Growth 알고리즘의 단점을 내제하고, 밀집도가 낮아 H-Main 알고리즘을 사용할 경우 H-Mine 알고리즘의 단점을 내제한다. 즉 2가지 종류의 알고리즘을 동시에 사용하므로 각각의 알고리즘을 사용할 정확한 기준 제시가 어렵고, 2가지 알고리즘의 단점을 내포한다.

따라서 본 논문에서는 단일 알고리즘을 사용하여 밀집 빈발 항목 집합 및 희소 빈발 항목 집합 모두에 대해 작은 메모리 공간을 사용하면서도 효율적이고 빠르게 빈발 항목 집합을 탐색할 수 있는 CPFP-Tree라는 새로운 자료구조와 탐색 방법을 제안한다.

2. CPFP-Tree 알고리즘

본 논문에서 새롭게 제안하는 자료구조인 CPFP-Tree의 이해를 돕기 위해 대표적인 트리기반 알고리즘인 FP-Growth의 FP-Tree와 비교한다[4].

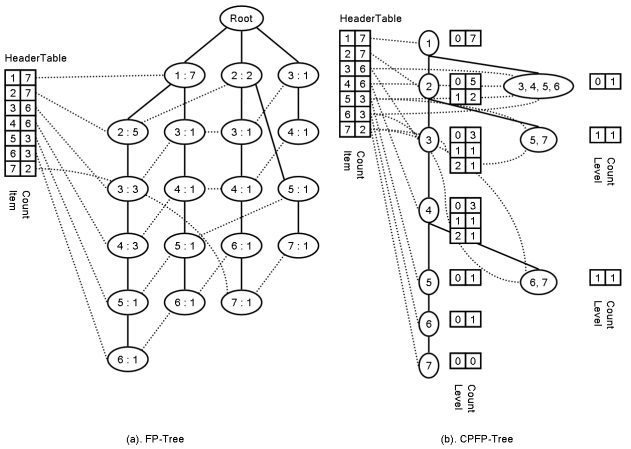

그림 2 FP-Tree와 CPFP-Tree의 비교

그림 2는 동일한 트랜잭션 데이터베이스로 구성된 것으로 이를 참조하여 살펴보면 CPFP-Tree는 FP-Tree에 비해 노드 수가 절반으로 감소했다. 노드 수의 감소는 트리의 크기를 줄여주며 이는 다시 빈발 항목 집합의 탐색 시간을 감소시킨다. 그리고 FP-Tree는 하나의 노드에 하나의 항목만을 가지고 있는 것에 반하여 CPFP-Tree는 하나의 노드에 여러 항목을 가질 수 있다. 이는 희소 빈발 항목 집합의 경우에 트리의 크기가 지나치게 커지고 가지의 수가 지나치게 많아지는 것을 방지한다. 또한 FP-Tree는 동일한 항목의 노드들의 연결을 각각의 노드들이 가지고 있지만 CPFP-Tree는 헤더 테이블이 모두 가지고 있다. 하나의 노드에 여러 항목을 가지고 있는 경우 분할이 발생할 수 있는데 이때 제안하는 자료구조는 노드들 간의 연결을 보다 쉽게 할 수 있다. 이는 트리의 구성 및 탐색 시간을 단축한다.

제안하는 자료구조를 이용하여 빈발 항목 집합을 탐색하는 방법은 원본 트랜잭션 데이터베이스 내의 모든 트랜잭션의 항목을 카운트하여 헤더 테이블을 구성하는 단계, 상기 구성된 헤더 테이블을 이용하여 CPFP-Tree를 구성하는 단계, 상기 구성된 CPFP-Tree 및 상기 헤더 테이블을 이용하여 각각의 항목에 대해 개별 CPFP-Tree를 구성하는 단계, 상기 구성된 개별 CPFP-Tree를 이용하여 빈발 항목 집합을 탐색하는 단계로 구성된다.

3. 결론

본 논문에서는 밀집 및 희소 빈발 항목 집합 모두에 효율적인 탐색 방법인 CPFP-Tree를 제안했고 이를 이용한 빈발 항목 집합 탐색 방법을 통해 밀집 및 희소 빈발 항목 집합 모두를 빠르게 탐색하는 효과가 기대된다. 따라서 빈발 항목 집합을 탐색하는 시스템 과 응용에서 활용이 가능하다. 또한 작은 메모리를 사용함으로써 제한된 시스템이나 대용량 데이터스트림의 빈발 항목 집합 탐색에도 효과적인 응용이 기대된다.

참고문헌

[1] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. Proceedings of the Data Mining and Knowledge Discovery, Springer Netherlands, 2007

[2] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules”, Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, 1994, pp. 487-499

[3] J.S.Park, M.S.Chen, and P.S.Yu, “An effective hash based algorithm for mining association rules”, Proceedings of the 1995 ACM SIGMOD international conference on Management of data, New York, NY, USA, 1995, pp. 175-186

[4] J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without Candidate Generation”, Proceedings of the ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 2000, pp. 1-12

[5] J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, and D. Yang, “H-Mine: Hyper-Structure Mining of Frequent Patterns in Large Databases”, Proceedings of the IEEE International Conference on Data Mining (ICDM), San Jose, California, USA, 2001, pp. 441-448

[6] J. Liu, Y. Pan, K. Wang, and J. Han, “Mining Frequent Item Sets by Opportunistic Projection”, Proceedings of ACM SIGKDD, Edmonton, Alberta, Canada, 2002

[7] A. Pietracaprina, and D. Zandolin, “Mining Frequent Itemsets using Patricia Tries”, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03), Melbourne, Florida, USA, 2003

[8] R. P. Gopalan and Y. G. Sucahyo, “High Performance Frequent Pattern Extraction using Compressed FPTrees”, Proceedings of the SIAM International Workshop on High Performance and Distributed Mining (HPDM), Orlando, USA, 2004