일반화된 연관규칙 발견을 위한 Level-based Data Mining 시스템

김온실[0] 박승수
이화여자대학교 컴퓨터학과
(onsil, sspark)@ewha.ac.kr

Level-based Data Mining System for Generalized Association Rules

On Sil Kim[0] Seung Soo Park
Dept. of Computer Science & Engineering, Ewha Womans University

요 약

데이터로부터 숨겨진 패턴을 추출하는 데이터마이닝 기법 중에서 연관규칙은 대용량의 데이터베이스에서 단위 트랜잭션 당 동시에 발생할 확률이 높은 항목들의 유형을 발견하는 기법이다. 연관규칙 탐사에서 개념계층(taxonomy)을 사용하여 보다 포괄적인 의미를 갖는 규칙을 찾아내는 연구가 일반화된 연관규칙이며 이를 통해 일반화 이전에는 간과될 수 있는 중요한 규칙을 발견할 수 있다.

일반화된 연관규칙에 관한 기존의 접근방법은 후보항목집합의 각 항목에 대한 개념계층상의 모든 조상들을 트랜잭션에 추가한 후 확장된 트랜잭션에 대해 지지도를 계산하는 방법이며, 이렇게 되면 연관규칙의 단점중의 하나인 계산량 문제가 더욱 두드러지게 된다.

이에 본 연구에서는 모든 개념계층 레벨이 아닌, 사용자가 관심 있는 레벨로 제한된 환경에서 연관규칙 탐사를 수행하여 규칙생성의 복잡도를 줄이는 시스템을 구현하였다. 그러나 모든 항목을 한 레벨로 일반화하는 데는 무리가 따르기 때문에 관심있는 항목의 경우 일반화 레벨을 따로 명시할 수 있도록 하여 사용자가 원하는 규칙을 발견하도록 하였다.

1. 서 론

폭발적으로 늘어나는 다양한 종류의 데이터를 수집하고 저장, 관리하는 데이터베이스 기술이 발전함에 따라 대량의 데이터 축적이 가능해졌다. 또한 데이터의 획득, 저장하는 능력에 비해 축적된 데이터를 분석하고 새로운 정보를 획득하는 능력이 뒤떨어짐에 따라 이를 해결하기 위한 기술로서 데이터마이닝이 대두되었다. 데이터마이닝은 대용량의 데이터에서 숨겨진 유용한 패턴을 추출하는 방법론이며, 실제로 여러 가지 데이터마이닝 기법이 제시되어 경영, 마케팅, 금융 등의 다양한 분야에서 활용되고 있다[1].

데이터로부터 숨겨진 패턴을 추출하는 연구중에서 연관규칙(Association Rules)은 Market Basket Analysis라고 불리기도 하며, 대용량의 데이터베이스에서 단위 트랜잭션 당 동시에 발행할 확률이 높은 항목들의 유형을 발견하는 기법이다[2][3]. 이러한 연관규칙은 고객 데이터베이스로부터 구매 품목들간의 관련성을 발견하여 교차판매 또는 상품 진열에 이용되거나, 우량고객에 대한 상품 카탈로그 발송의 Direct-mailing에 이용될 수 있다[3].

연관규칙 탐사에서 개념계층(taxonomy)을 사용하여 보다 포괄적인 의미를 갖는 규칙을 찾아내는 연구가 수행되었는데, 이를 일반화된 연관규칙(Generalized Association Rules)이라 한다[4]. 트랜잭션의 각 항목이 속하는 계층적 카테고리인 개념계층을 연관규칙 탐사에 이용하여 보다 상위 레벨에 해당하는 일반적인 규칙을 찾는 것이다.

일반화된 연관규칙에 관한 기본적인 접근법은 Level-Cross로서, 서로 다른 개념계층 사이의 연관규칙을 찾는 방법이다[3][4][5]. 이러한 접근 방법은 연관규칙의 단점인 계산량을 고려할 때 바람직하지 않을 수 있다. 이에 본 논문에서는 일괄적인 일반화를 제시하고 특정 아이템의 경우 일반화 레벨을 따로 명시하여 예외를 허용할 수 있는 시스템을 구현하였다.

본 논문의 구성은 다음과 같다. 2장에서는 일반화된 연관규칙의 개념 및 필요성과 기존의 접근방법에 대해 간단히 살펴보고, 3장에서는 일괄적인 일반화와 특정 아이템의 경우 별도의 일반화 레벨을 명시할 수 있는 시스템의 구성을 보인다. 4장에서 구현된 시스템을 이용한 실험과 결과를 보이며, 5장에서 결론을 내린다.

2. 관련연구

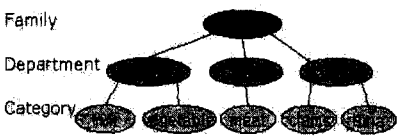
2.1 Generalized Association Rules

일반적으로 물품구매정보에 나타나는 각 항목은 상품 코드로 구성되며, 각 상품이 속하는 계층적 카테고

* 본 연구는 KOSEF(98-0102-01-01-3)의 지원에 의해서 수행되었음

리인 개념계층(taxonomy)이 함께 제공된다[4]. 이러한 개념계층의 하위레벨로 내려갈수록 최소 지지도를 만족하지 못하는 규칙이 생성되는 경향이 있으며, 그렇기 때문에 중요한 규칙이 간과될 수 있다[1][4]. [그림 1]와 같은 구매내역과 개념계층 정보에서 다음의 두 가지 규칙을 생각해 보자.

- Category level : fruit -> clams (support = 1)
- Department level : produce->canned (support = 3)



Transaction	Items
1	fruit, tuna
2	meat, tuna
3	fruit,meat,clams
4	vegetable,clams

min. support = 2

[그림 1] taxonomy 정보와 물품구매정보

개념계층 상에서 하위 레벨인 첫번째 규칙의 지지도는 최소 지지도(=2) 보다 작은 값을 갖기 때문에 규칙이 될 수 없으나, 이를 Department 레벨로 일반화시킨 두 번째의 경우 규칙으로 채택될 수 있음을 알 수 있다. 그렇기 때문에 항목들을 좀 더 포괄적인 범위로 일반화시켜, 간과될 수 있는 규칙을 발견하기 위해서 개념계층을 이용한 일반화가 필요하게 된다.

일반화된 연관규칙에 관한 기존의 접근방법[3][4][5]은 Level-cross Association Rules로서, 후보항목집합의 각 항목에 대한 개념계층상의 모든 조상들을 트랜잭션에 추가하여 확장된 트랜잭션에 대해 지지도를 계산하는 방법이다. 이렇게 되면 연관규칙의 단점중의 하나인 계산량 문제가 더욱 두드러지게 된다. 각 항목들의 발생빈도를 계산하기 위한 데이터베이스 스캔은 여러 단계에 걸쳐 필요하며 항목들이 증가할수록 계산량은 기하급수적으로 늘어나게 되는 것이다.

이에 본 연구에서는 모든 개념계층 레벨이 아닌, 사용자가 관심 있는 레벨로 제한된 환경에서 연관규칙 탐사를 수행하여 규칙생성의 복잡도를 줄였다. 그러나 모든 항목을 한 레벨로 일반화하는 데는 무리가 따르기 때문에 관심있는 항목의 경우 일반화 레벨을 따로 명시할 수 있도록 하여 사용자가 원하는 규칙을 발견하도록 하였다.

3. 시스템 설계 및 구현

시스템의 주된 기능은 다음과 같으며 각 기능은 독립적으로 사용될 수도 있으며 함께 사용될 수도 있다.

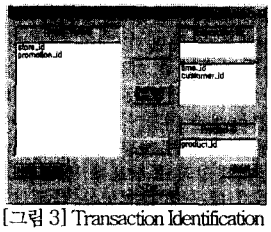
- 일반화(Generalization) : 개념계층을 통한 일반화로, 간과될 수 있는 규칙의 발견과 규칙 생성의 복잡도 줄임
- 특정 항목 선정(Specific item) : 일괄적인 일반화 레벨에서 벗어나는 특정 항목의 선정으로 별도의 일반화 레벨을 명시

시스템은 [그림 2]의 4가지 모듈로 구성되며 그 순서는 시스템의 작업 흐름과 같다.

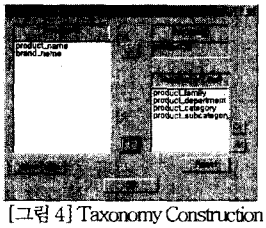
[그림 2] 시스템 구성도

단계별 작업으로 우선은 분석할 데이터베이스를 결정, 거래내역 정보를 갖는 트랜잭션 테이블과 개념계층 정보를 갖는 개념계층 테이블을 선택하게 된다. 그 후, Transaction identification 모듈을 거쳐 거래내역으로부터 트랜잭션을 인식한다[그림 3]. 이때 해당 테이블에 트랜잭션 ID가 부여되어 있다면 해당 필드를 선택하고 없다면 ID를 유도할 수 있는 필드들을 선택하게 된다.

Taxonomy construction 모듈에서는 분석할 개념계층의 레벨들을 선택, 적절한 계층을 구성하고 이를 조정하게 된다[그림 4].

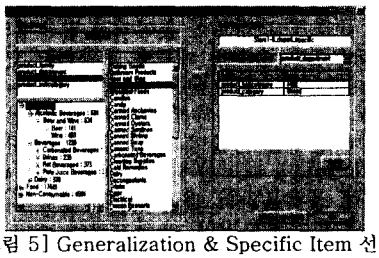


[그림 3] Transaction Identification



[그림 4] Taxonomy Construction

다음 단계가 시스템의 핵심 기능을 담당하는 Generalization & Specific Item 선정 모듈이며, 전체적인 일반화 레벨의 결정과 특정 항목의 일반화 레벨의 선정이 이루어진다[그림 5]. 개념계층의 각 레벨을 선택하면 각 레벨에서 나타나는 모든 항목들을 볼 수 있으며 특정 항목과 이의 일반화 레벨을 정함으로 특정 항목 선정이 이루어진다. 왼쪽 하단 부분에는 Taxonomy construction 모듈을 거쳐 생성된 트리 구조의 개념계층 정보가 각 항목의 빈도수와 함께 표시되어 사용자의 결정에 도움을 주고 있다.



[그림 5] Generalization & Specific Item 선정

마지막이 위의 과정을 거쳐 변환된 트랜잭션을 대상으로 연관규칙을 찾는 단계이며 본 논문에서는 범용 데이터마이닝 툴인 Clementine(SPSS社)을 사용하였다[6].

4. Experiments

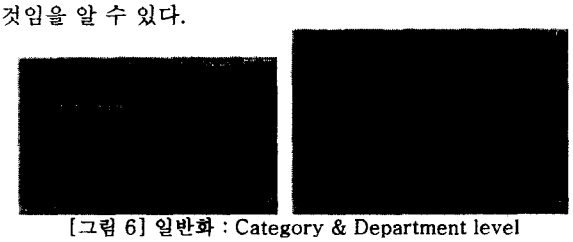
위의 시스템을 적용하여 몇 가지 실험을 하였다. 사용 데이터는 MS SQL Server의 샘플 데이터로, 251,395건의 개별적인 거래내역을 갖고 있으며 실험에서는 이중 일부인 24,278건을 사용하였다. 데이터의 개념계층(taxonomy hierarchy)을 살펴보면, 최상위인 product_family 레벨은 3개의 항목이 있고, department 레벨은 21개, category 레벨은 47개, subcategory 레벨은 110개, product_name은 1,560개의 항목이 있다.

2%의 최소지지도와 60%의 신뢰도를 기준으로 아래의 [표 1]과 같이 경우를 나누어 연관규칙을 찾아보았다.

[표 1] 전체적인 일반화 레벨 및 특정 항목의 일반화 레벨 선정

	적용 기능	일반화 level
Case 1	Only Generalization	Subcategory level
Case 2		Category level
Case 3		Department level
Case 4	Generalization	Department level
	Specific Items	Bread : Category level Milk : Subcategory level
Case 5	Only Specific Items	Drink : Family level Baked Foods : Department Snack Foods : Category Eggs : Category Magazines : Category

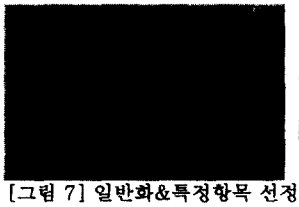
실험 결과 일반화 레벨이 가장 낮은 Case 1의 경우 아무런 규칙도 생성되지 않았으며, Case 2의 경우 10개의 규칙이 발견되었고 Case 3의 경우 18개의 규칙이 발견되었다[그림 6]. 이는 하위 레벨에서 threshold를 만족하지 못해 버려졌던 규칙이 상위 레벨로의 일반화를 통해 발견된 것임을 알 수 있다.



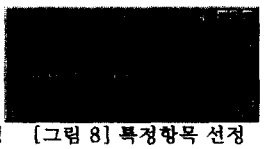
[그림 6] 일반화 : Category & Department level

Case 4의 경우 17개의 규칙이 발견되었으며 별도의 레벨로 선택했던 항목들이 규칙에서 나타났다[그림 7].

Case 5의 경우 선택했던 항목들로 구성된 2개의 규칙이 발견되었다[그림 8].



[그림 7] 일반화&특정항목 선정



[그림 8] 특정항목 선정

위의 5가지 실험을 종합해 볼 때, 일반화 레벨이 높으면 높을수록 threshold를 만족시켜 더 많은 규칙이 발견되는 경향을 보이며, 특정 항목 선정시 해당 항목이 나타나는 규칙이 발견되어 사용자가 원하는 형태의 규칙이 생성됨을 확인하였다.

5. 결론 및 향후연구

마이닝 결과 생성된 규칙의 흥미도 측정 지표로서 완전성(Completeness)과 최적성(Optimization)이 있다.

완전성은 모든 규칙을 찾기 위한 작업수행을 말하는데 반해, 최적성은 특정 영역을 한정해서 관심있는 규칙만을 찾는 것이다.

연관규칙은 특히 계산량이 문제시되는 마이닝 기법으로 항목이 증가할수록 계산량은 기하급수적으로 늘어나게 된다. 그렇기 때문에 개념계층을 통한 일반화가 필요하며, 최적성 접근방법으로 일괄적인 일반화를 구현하는 시스템을 설계하고 구현하였다. 또한 모든 항목을 한 레벨로 일반화하는 데는 무리가 따르기 때문에 관심 있는 항목의 경우 일반화 레벨을 따로 명시할 수 있도록 하여 사용자가 원하는 규칙을 발견하도록 하였다.

앞으로는 규칙의 표현시 각 항목이 속하는 개념계층 레벨의 병행표기가 필요하며 특정 항목 선정시 도움을 줄 수 있는 통계분석이 추가되어야 하겠다.

6. 참고문헌

[1] Michael J. A. Berry and Gordon Linoff, "Data Mining Techniques for Marketing, Sales and Customer Support", Wiley Computer Publishing, 1997

[2] R. Agrawal and R. Srikant, "Fast algorithms for Mning association rules", Proceedings of the 20th VLDB Conference, 1994

[3] Jochen Hipp, Ulrich Guntzer and Cholamreza Nakhaeizadeh, "Algorithms for Association Rule Mining - A General Survey and Comparison", SIGKDD Explorations, 2(1) p:58-64, 2000

[4] R. Srikant and R. Agrawal, "Mining Generalized Association Rules", Proceedings of the 21th VLDB Conference, Zurich, Swizerland, 1995

[5] Shiby Thomas and Sunita Sarawagi, "Mining Generalized Association Rules and Sequential Pattern Using SQL Queries", Proceedings of the 4th Intl. Conf. On Knowledge Discovery and Data Mining(KDD'98) p:344-348, 1998

[6] Clementine User Guide Version 5