데이터 마이닝을 위한 연관규칙의 다중 값 속성 처리방법

김산성°, 김명원
숭실대학교 컴퓨터학부
kbibboss@orgio.net, mkim@comp.ssu.ac.kr

Processing Multi-Valued Attributes in Association Rules for Data Mining

San-Sung Kim°, Myoung-Won Kim
School of Computing, Soonsil University

요 약

다중 값이란 속성 값이 집합인 것을 말한다. 즉, 관계형 데이터베이스에서 자료 유형이 집합인 속성을 의미한다. 이러한 다중 값 속성 처리는 기존 데이터마이닝 기술 자체로는 처리할 수 없으며 후처리나 전처리 과정을 이용하여 처리하고 있다. 전처리나 후처리 과정을 통해 처리할 경우 수행과정에 있어 많은 시간이 소요되고 혹은 타당하지 않은 규칙이 생성되는 문제점을 가지고 있다. 특히 연관화 기법 특성상 분석하고자 할 항목이 증가할수록 연관성의 수가 지수(exponential)단위이기 때문에 이를 해결하는데는 상당한 어려움이 따르게 된다. 본 논문에서는 관계형 데이터베이스 테이블 구조에서 데이터 마이닝의 수행을 위한 전처리나 후처리의 과정을 고려하지 않음으로 위에서 언급된 문제점들을 해결하고자 한다. 특히 데이터 변환 작업 없이 정량적 (Quantitative)연관 규칙과 연관 규칙(Market Basket Analysis)의 혼합 형태의 규칙을 생성할 수 있게끔 알고리즘을 확장하여 보다 효율적인 규칙이 생성될 수 있도록 한다. 마지막으로 Each Movie 데이터를 사용하여 확장한 알고리즘의 다중 값 속성 처리 방법의 효율성과 타당성을 검증한다.

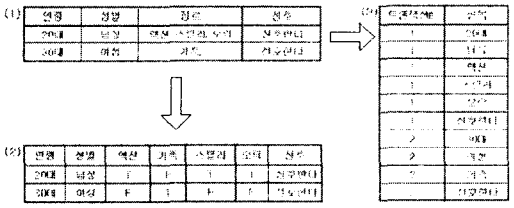
1. 서 론

많은 웹 사이트나 서비스 업체에서는 다양한 고객의 요구를 만족시켜주기 위한 규칙을 생성하기 위해 고객에게 많은 정보를 요구하고 있다. 예를 들면, 웹 사이트 회원가입 시 입력하는 정보로 하나 이상의 값을 가질 수 있는 취미나 관심분야와 같은 속성정보들을 말한다. 이러한 여러 가지 속성 정보들은 고객에게 요구됨과 동시에 효율적으로 처리할 수 있는 방법이 필요하게 되었다. 따라서 본 논문에서는 [그림 1]처럼 표현된 취미속성을 다중 값 속성이라고 정의한다. 즉, 속성 값이 집합인 속성이다.

사용자	성별	직업	취미	컨텐츠
사용자1	남성	학생	여행, 낚시, 등산	제주도
사용자2	남성	학생	여행, 낚시	제주도
사용자3	여성	학생	등산,	설악산
사용자4	남성	회사원	영화감상, 음악감상	영화

[그림 1] 다중 값 속성

다중 값 속성은 속성과 속성 값이 일대다의 관계를 갖게 되고 다중 값의 속성이 포함된 데이터를 마이닝 하기 위해선 전처리(Preprocess) 단계에서 일대일의 관계로 변환해 주어야 한다. 변환 방법은 [그림 2]의 (2)와 같이 다중 값 속성의 값 개수만큼 속성 개수를 늘리는 테이블의(tabular) 입력형식과 (3)과 같은 레코드 개수를 늘리는 트랜잭션의(transactional)입력형식으로의 변환 방법이 있다. 이러한 두가지 입력형식에서 속성개수를 늘리는 테이블의 입력형식은 고려해야할 항목수가 (다중값 개수*2)로 늘어나는 문제점이 발생하며, 레코드개수를 늘리는 트랜잭션의 입력형식은 빈발항목집합 조합의 크기가 늘어나는 문제점이 발생한다. 또한 이러한 입력형식의 결과인 규칙은 후처리에 통한 필터링(filtering) 절차를 필요로 하게 되며, 더욱이 마이닝 기법과 데이터의 특성에 대한 사전 지식이 없는 관리자에게는 마이닝을 수행하기 위한 전처리단계와 후처리단계가 많은 부담이 되고 있다. 따라서 본 논문에서는[1] 데이터 마이닝 기법 중에 Apriori 연관화 기법에서의 다중 값 속성 처리 방법을 제안하여 효율적으로 규칙을 생성한다. 본 논문에서 제안하는 방법은 입력 자료구조에 벡터(vecter)를 사용함으로 다중 값의 개수에 따라 유동적으로 벡터 크기를 할당하여 메모리를 효율적으로 사용하고 항목에 주어지는 인덱스를 이용하여 빈발항목집합을 구한다. 따라서 적은 메모리의 사용과 마이닝 단계 축소를 통해 규칙 생성 시간을 단축시키며 정량적(quantitative)연관 규칙과 MBA (Market Basket Analysis) 연관 규칙의 혼합 형태의 규칙을 생성한다.



[그림 2] 다중 값 속성처리를 위한 자료 입력형식

본 논문의 구성은 다음과 같다. 2장에서는 데이터마이닝에 널리 사용되는 기법인 Apriori 연관화 알고리즘과 기존의 마이닝 툴인 SPSS의 Clementine, SAS의 Enterprise Miner, IBM의 Intelligent Miner에서의 연관규칙을 뽑기위한 자료 입력형식에 대해 설명한다. 3장에서는 제안하는 다중 값 속성 처리방법에 대해 설명하며, 4장에서는 제안한 다중 값 속성 처리방법에 대해 EachMovie 데이터를 통한 실험결과를 보인다. 마지막으로 5장에서는 논문의 결과와 향후연구를 검토한다.

2. 관련연구

2.1 연관규칙[1][2][3][4]

연관 규칙이란 하나의 거래(transaction)나 사건에 포함되어 있는 항목(item)들의 경향을 파악해서 상호 연관성을 발견하는 것이다. 데이터베이스 내에서 연관 규칙을 찾아내는 것은 데이터 마이닝의 측면에서는 매우 중요한 문제이다. 트랜잭션의 집합이 주어질 때, 각각의 거래는 항목의 집합이다. 연관성 규칙을 계산 할 수 있는 방법으로는 지지도(Support), 신뢰도(Confidence)가 있다.

지지도란 전체 거래 중에서 항목 X와 항목 Y를 동시에 포함하는 거래가 얼마나 있는가 하는 것이다.

$$S = P(X \cap Y) = \frac{\text{항목}X\text{와 항목}Y\text{를 포함하는 거래의수}}{\text{전체거래수}N}$$

1) 본 연구는 한국과학기술원 뇌신경정보학연구사업의 지원에 의하여 수행되었습니다.

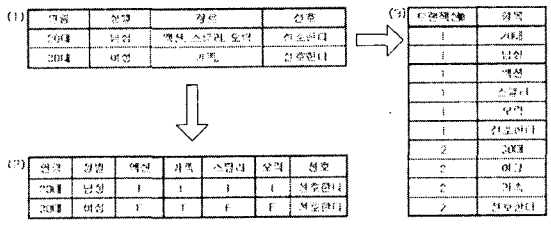
신뢰도는 항목 X를 포함하는 거래 중에서 항목 Y가 포함될 확률은 어느 정도인가를 의미한다.

$$C = P(X|Y) = \frac{P(X \cap Y)}{P(X)} = \frac{\text{항목}X\text{와 항목}Y\text{를 포함하는 거래수}}{\text{항목}X\text{를 포함한거래수}}$$

일반적으로 연관규칙의 표현에는 정량적(quantitative) 연관규칙과 MBA(Market Basket Analysis)연관규칙으로 구분할 수 있다. 두 연관규칙은 단지 항목을 정의하는 데에 차이가 있다. 즉, 정량적 연관규칙은 항목을 (속성, 속성값)으로 정의하며 하나의 거래에서 발생하는 항목은 임의의 속성에 속성값이 임의의 값을 가지고 알고리즘을 수행하는 반면 MBA 연관규칙은 특정 속성의 속성값을 항목으로 정의하며 하나의 거래의 특정 속성은 여러 가지 값을 가지고 알고리즘을 수행한다. 즉, 생성되는 정량적 연관규칙은 조건부에 같은 속성이 두 번이상 빈발항목 집합에 포함될 수 없고 MBA 연관규칙은 빈발항목 집합에 같은 속성이 한번 이상 포함될 수 있다. 단 MBA 연관규칙은 특정 속성만을 이용한다.

2.2 기존 툴(Tool)에서의 연관규칙 입력형식

기존에 가장 보편적으로 많이 사용하는 툴인 SPSS의 Clementine, SAS의 Enterprise Miner, IBM의 Intelligent Miner의 연관규칙에서 다중 값 속성을 처리하기 위한 데이터 입력형식은 [그림 3]과 같다. [그림 3]의 (1)에서 '장르'라는 다중 값 속성을 처리하기 위해 (2)의 테이블의(tabular) 입력형식과 (3)의 트랜잭션의(transactional) 입력형식으로 변환해야 한다. (2)의 입력형식은 장르가 갖는 속성 값 개수만큼 속성수를 늘려줌으로 항목개수가 늘어나는 문제점이 있고 (3)의 입력형식은 모든 속성값을 개별적인 레코드로 변환해야 하며 빈발항목집합의 조합의 크기가 커지는 문제점을 갖게 된다. 결과적으로 다중 값 속성을 처리하기 위한 기존 툴의 입력형식은 전처리 과정을 수행해야 하며 전처리에 따른 수행시간을 소요하게 된다. SPSS의 Clementine는 (2), (3)의 입력형식을 모두 지원하고 있으며, SAS의 Enterprise Miner, IBM의 Intelligent Miner 경우에는 (3)의 입력형식만을 지원하고 있다.



[그림 3] 기존 툴에서의 자료입력 형식

2.2.1 테이블(tabular)의 입력형식

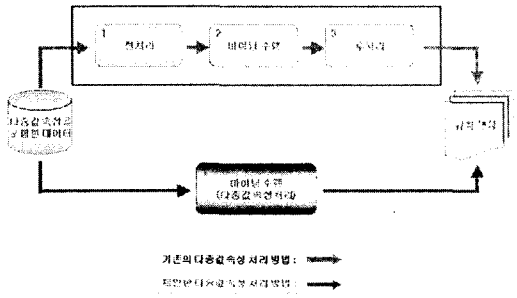
[그림 3]에서 (2)는 테이블의 입력형식으로 변환한 것이다. 즉 [그림 3]에서 (1)의 장르 속성이 가지는 속성 값인 액션, 스릴러, 오락, 가족 각각을 하나의 속성으로 표현한다. 각각의 속성은 T과 F 값을 갖는다. F의 값은 선택하지 않은 값으로 사용자가 좋아하는지에 대한 선호를 알 수가 없으며 T의 값은 선택한 속성 즉, 좋아하는 취미를 표현한다. 늘어난 속성은 Boolean값을 갖게 되고 (다중값 개수*2)개의 항목이 증가하게 되어 마이닝 수행시 많은 계산량을 요구하게 된다. 생성되는 규칙에서는 속성 값이 F인 액션, 오락등의 속성이 빈발항목으로 포함되어 있다. F라는 값은 고객 선호도를 알 수 없는 값으로 이러한 값이 규칙생성에 포함된다면 규칙자체의 정확도를 떨어뜨릴 수밖에 없다. 또한 F라는 값을 포함시키지 않기 위해서는 전처리나 후처리를 통한 추가작업이 필요하며 다중 값 개수만큼 속성개수가 늘어나므로 고객정보를 테이블 구조 또한 고정적이고 비효율적이 된다.

2.2.2 트랜잭션(transactional)의 입력형식

[그림 3]에서 (3)은 트랜잭션의 입력형식으로 변환한 것이다. 이러한 입력 형식은 트랜잭션 ID와 항목으로 구분되며 항목은 단지 하나의 값만을 가진다. 따라서 다중 값 속성을 포함한 모든 속성 값이 개별적인 레코드로 변환되는 것이다. [그림 3]의 (1)에서 첫 번째 레코드가 갖는 장르속성 값인 액션, 스릴러, 오락과 연령, 성별, 선호의 값을 6개의 레코드로 변환하여 표현된다. 이러한 입력형식은 모든 속성 값의 개수만큼 레코드가 늘어나므로 비효율적으로 시스템 메모리를 사용하게 되고 빈발항목집합 조합의 크기가 커지게 되어 마이닝 수행 시 많은 시간을 요구하게 되는 문제점이 발생한다. 또한 대용량 데이터를 처리하기 위한 트랜잭션의 입력형식은 정량적(Quantitative) 연관 규칙과 MBA(Market Basket Analysis) 연관 규칙의 혼합 형태의 규칙이 생성시키기에는 비효율적이다.

3. 다중값 속성 처리방법

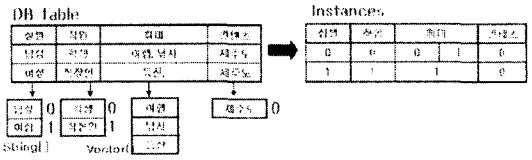
다중 값이란 속성 값이 집합인 것이다. 즉, 관계형 데이터베이스에서 자료 유형이 집합인 속성을 말한다. 예를 들면, 취미, 관련분야, 컨텐츠 키워드, 영화장르와 같은 속성들을 의미한다. 즉, 사용자가 가질 수 있는 취미나 관심분야는 하나 이상일 수 있고 또한 컨텐츠에 포함된 키워드 역시 하나 이상일 수 있다.



[그림 4] 다중 값 속성 처리 절차

다중 값 속성에 대한 기존의 처리방법은 [그림 4]에서 1,2,3처럼 전처리와 후처리 등의 추가적인 작업이 필요하다. 본 논문에서는 기존 처리방법의 문제점을 분석하고 추가적인 작업절차 없이 다중 값 속성을 처리하는 방법을 제안한다.

3.1 처리방법



[그림 5] 다중값 속성 처리방법의 자료구조

관계형 데이터베이스에 저장되어 있는 자료형태를 그대로 입력으로 받아들여 처리할 수 있도록 속성 값에 할당되는 인덱스를 이용하여 아래 [그림 5]와 같이 Item이라는 빈발항목집합 클래스를 생성하게 된다. Item 클래스는 빈발항목을 표현하는 int형 타입의 m_item배열 변수와 항목의 빈도수를 나타내는 int형 타입의 m_counter변수를 가지고 있다. Item객체 생성시에는 m_item배열 변수를 -1로 초기화 한다. 그리고 빈발항목들의 인덱스를 각각 m_item배열에 표시한다. 여기서 -1값은 빈발항목 집합에서 고려하지 않는 항목을 표현하며 다중 값 속성의 각각의 다중 값도 하나의 속성으로 처리하여 m_item배열에 표현된다. 특히 다중 값 속성은 vector를 사용하여 유동적으로 배열의 크기를 조작하여 시스템 메모리 낭비를 최소화한다.

4. 실험

4.1. 실험데이터

실험 데이터는 Each Movie데이터로 총 사용자가 72916명이고, 영화는 1628편, 장르는 *액션, 애니메이션, 외국예술, 고전, 코미디, 드라마,*

*가족, 공포, 로맨스, 스릴러*의 10가지로 분류되어 있다. 이러한 세 개의 테이블을 조인(Join)시켜 하나의 실험테이블로 재구성하고 영화 항목 테이블의 장르 속성이 다중 값 속성이다.

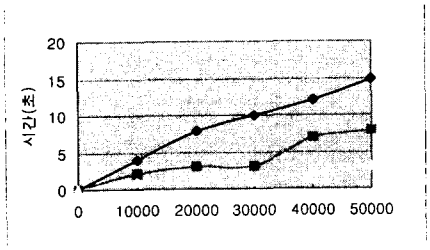
4.2 실험방법

조인된 실험테이블은 166,583개의 트랜잭션을 갖는다. 조인 시에 사용된 속성으로는 연령, 성별, 장르(10개의 다중값), 선호도 속성을 사용하였다. 장르는 하나이상의 값을 가질 수 있으며 실험은 장르 속성값 개수와 트랜잭션 개수를 늘려가며 마이닝 기법중 연관기법 알고리즘 수행시간을 비교 분석했다. 알고리즘 파라미터 값으로는 최소지지도 0.1, 최소신뢰도 0.5로 주었으며 다중 값 개수를 5개, 10로, 트랜잭션 개수를 10,000개 단위로 10,000개에서 50,000개로 늘려가며 실험을 하였다.

4.3 실험결과

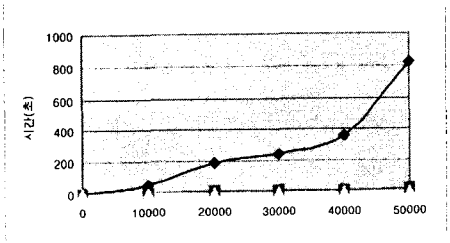
전처리 수행을 통해 변환된 두 가지 입력방식과 제안한 테이블 형식의 트랜잭션별 마이닝 수행시간을 비교 하였다. [그림 6]는 다중 값 개수가 5개일 때의 마이닝 수행 시간 결과로 제안한 테이블 입력형식이 가장 적은 수행 시간이 요구됨을 알 수 있다.

◆:테이블입력형식 ■:트랜잭션입력형식 ▲:제안한입력형식



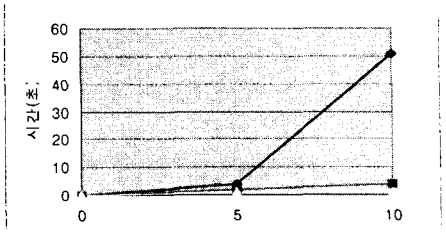
[그림 6]트랜잭션별-다중값개수 5개일때

[그림 7]에서 마이닝 수행 시간이 가장 많이 소요한 테이블의 입력형식은 항목개수에 비해 (값개수 5 * 2)만큼의 항목개수가 증가하게 됨으로 연관성의 수가 지수(exponential) 단위로 증가한다. 즉, 마이닝 수행시간이 지수적으로 증가하게 되는 것이다. 결과적으로 트랜잭션 입력형식의 빈발항목집합 조합의 크기보다 테이블의 입력형식에서 증가된 항목개수가 마이닝 수행시간에 크게 영향을 미치는 것을 알 수 있다.



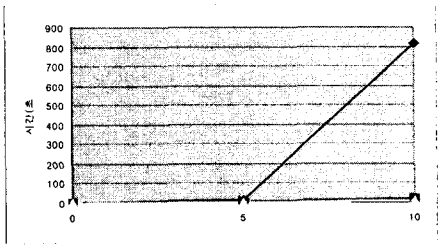
[그림 7]트랜잭션별-다중값개수 10개일때

[그림 8]은 트랜잭션이 10,000개일때의 다중 값 개수를 5, 10개로 늘려가며 마이닝 수행에 소요되는 시간 변화 분석한 결과다. 다중 값 개수가 5개에서 10개로 늘어남으로써 항목개수가 (다중값개수 *2)로 증가하는 테이블의 입력형식이 52초로 가장 길게 마이닝 수행 시간을 소요한다.



[그림 8]다중값개수별-다중값개수 5개일때

[그림 9]는 트랜잭션이 50,000개일때의 다중 값 개수를 5, 10개로 늘려가며 마이닝 수행에 소요되는 시간을 분석한 결과다. 트랜잭의 개수가 [그림 8]에 비해 5배로 증가할 때 테이블의 입력형식의 마이닝 수행 시간은 820초가 된다. [그림 9]의 결과는 처리해야 할 데이터가 대용량일수록 다중 값 속성 처리는 비효율적임을 알 수 있다.



[그림 9]다중값개수별-다중값개수 10개일때

기존의 입력형식으로는 다중 값 속성을 처리하기에는 많은 전처리 수행시간과 연관규칙을 찾는 마이닝 수행시간을 크게 요구하는 것을 알 수 있다. 위의 결과를 통해서 제안한 테이블 입력형식이 다중 값 속성을 가장 효율적으로 처리하고 있음을 보여 준다.

6. 결론 및 향후연구

본 논문에서는 관계형 데이터베이스에서 테이블을 구성하는 여러 속성 중에 하나의 속성이 여러 가지 값을 가질 수 있는 다중 값 속성을 정의하고 이러한 다중 값 속성을 처리할 수 있는 방법을 제안하였다. 연관기법에서의 다중 값 속성 처리를 위한 기존의 두가지 입력 형식, 즉 트랜잭션의 입력형식과 테이블의 입력형식의 문제점은 전처리 수행의 시간 소요와 마이닝 수행 시의 시스템 사용 메모리와 연관 규칙을 얻는데 필요한 시간이 크게 증가한다는 것이다. 본 논문에서 이와 같은 문제를 해결하고자 입력형식에 대한 자료구조를 제안하여 다중 값 개수가 많아지거나 처리해야 할 데이터가 늘어날수록 효율적으로 처리됨을 실험을 통해 보였다. 특히 일반 항목속성 값과 인구통계학적(Demographic)속성 값을 연관성을 추출하여 보다 다양하고 효율적인 규칙 적용을 위한 정량적 연관규칙과 MBA연관규칙의 혼합형태의 규칙을 생성하였다. 향후 연구로는 데이터 마이닝의 여러 기법에서의 다중 값 속성 처리를 위해 우선적으로, 군집화 기법에서의 다중 값 속성 처리 방법과 적절한 평가방법에 대한 연구가 필요하다.

참고문헌

[1] Ramakrishnan Srikant, Rakesh Agrawal, "Mining Generalized Association Rules", VLDB, p 407-419, 1995

[2] Ramakrishnan Srikant and Rakesh Agrawal, "Mining quantitative association rules in Large Relational Tables", In Proceedings of the ACM SIGMOD Conference on Management of Data, p 3-8, June 1996.

[3] R. Agrawal and R. Stikant, "Fast algorithms for mining Association rules", In Proceedings of the 20th VLDB Conference, pp487-499 Santiago, Chile, Sept., 1994

[4] Rakesh Agrawal, Tomasz Imielinski, Arun Swami, "Mining Association rules between sets of items in large database", In Proc. of the ACM SIGMOD Conference on Management of Data, pages 207-216, 1996