설문 데이터를 위한 다차원 연관 규칙 마이닝

Multi-Dimensional Association Rule Mining in Survey Data

이 정 수[a], 김 교 정[b]

[a] 숙명여자대학교 자연과학연구소

서울 특별시 용산구 청파동 2가 53-12, 140-742

Tel : +82-2-710-9379, Fax : +82-2-710-9296, E-mail : integer@sookmyung.ac.kr

[b] 숙명여자대학교 정보과학부 멀티미디어학과

서울 특별시 용산구 청파동 2가 53-12, 140-742

Tel : +82-2-710-9428, Fax : +82-2-710-9296, E-mail : kiochkim@sookmyung.ac.kr

Abstract

본 논문에서는 인문 사회과학 분야의 방대한 설문 데이터를 처리하기 위해 기존의 설문 항목들간의 평면적 관계에만 국한 되었던 연구에 대해 설문데이터 다차원 연관규칙 마이닝 시스템을 설계하고 데이터 간의 연관규칙을 탐사한다. 즉, 직관적으로 분류될 수 있는 기준에 따라 클러스터링을 실행하여 데이터를 분류한 후 각 클러스터로부터 다차원 연관 규칙을 탐사하는 시스템을 제안함으로써 보다 강력한 연관규칙을 탐사한다.

Keywords: **연관규칙 (association Rule), 데이터 마이닝 (data mining), 다차원 연관 규칙 (multi dimensional association rule)**

1. 서론

최근 사회 각 분야에서 수많은 설문 조사를 통하여 특정 관심 분야에 관한 여론을 파악하는 것을 볼 수 있다. 이러한 자료의 급증에도 불구하고 특정 항목을 선정하여 그 관련성을 파악해야 하는 통계 처리로는 수집되어진 설문 자료의 내재된 정보를 충분히 활용하지 못하고 있다.

이러한 정보 속의 정보를 데이터 마이닝 기법을 이용하여 신뢰성 있는 규칙을 탐사하고 탐사 되어진 규칙을 통하여 데이터베이스 내에 내재되어있는 정보를 알아낼 수 있다. 또한 설문 조사의 특성상 유사 질의항목을 동일 범주로 묶어 줌으로써 응답자의 편의를 도모할 수 있다. 이에 본 논문[1]에서는 숙명여자대학교 아태여성정보통신센터의 아태지역 6개국의 여성 정보화 현황조사에서 Scientific Sampling Methodology에 의해 획득한 데이터 중 필리핀 마닐라에 서 채취한 1000개의 데이터를 사용한다.

제공되어진 데이터를 통하여 여성 정보화와 밀접한 관련이 있는 질의 문항을 찾아내는데 항목간의 연관성을 파악하고자 연관 규칙(association rule) 을 사용하였으며, 가시적인 분류를 기준으로 일차원 혹은 평면 상의 규칙이 아닌 다차원에서의 연관규칙[1]을 탐

[1] 본 연구는 2002년 숙명여자대학교 교비 연구과제의 지원으로 수행된 것임.

사함으로써 보다 정확한 규칙을 사용자에게 제공한다.

본 논문의 구성은 2장에서는 설문 데이터를 위한 데이터 마이닝 시스템에 대해 살펴보고 3장에서는 본 연구에서 제안한 다차원 연관 규칙 마이닝 시스템에 관하여 살펴보도록 한다. 그리고 4장에서 결론을, 5장에서는 참고문헌을 소개하도록 한다.

2. 설문데이터를 위한 다차원 연관 규칙 마이닝

기존의 설문 데이터를 마이닝 하는 시스템의 경우 통계학 분야에서는 전문 지식인의 도움을 빌어 특정 항목들 간의 함수관계를 파악하는 반면, 전산학에서의 마이닝 연구는 제시되어진 최소 지지도(min. support)와 최소 신뢰도(min. confidence)를 바탕으로 모든 항목들간의 빈도를 계산함으로써 각 항목간의 연관성을 파악한다. 이러한 일 차원 선상의 연관 규칙은 정량적인 항목 값의 처리를 불가능하게 만든다.

본 논문에서 제안하는 마이닝 시스템은 정량적인 데이터를 처리하기 위하여 설문데이터의 가시적인 분류 항목들을 기준으로 다차원 연관 규칙을 탐사한다.

다음 그림 1은 본 논문에서 제시한 마이닝 시스템인 CARE(Customized Association Rule Engine)의 구조도 이다.

본 시스템의 동작은 다음과 같다.

- Data Preprocessing Engine에서는 주어진 레코드 데이터를 본 시스템의 형식에 맞게 데이터 전처리 엔진을 통하여 변환한다. 예를 들어, 연령의 경우 응답자의 실제 연령을 범주형 데이터로 변환해 주는 등 의 전처리 과정을 통하여 본 시스템에 맞는 데이터로 변환한다.
- Clustering Engine 에서는 제시된 최소 지지도(min. support)를 바탕으로 가시적으로 분류할 수 있는 항목으로 설문 항목의 분류한 후 각 분류된 항목에 맞게 유동적인 지지 빈도수를 적용하여 클러스터를 분류한다.

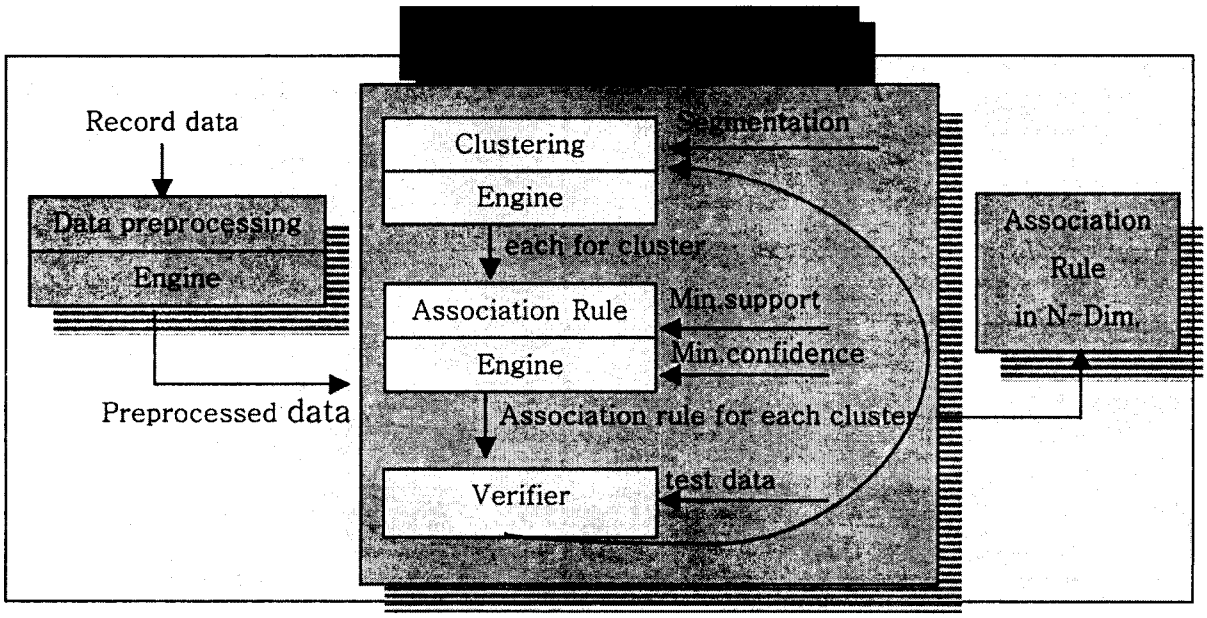

그림 1 Customize Association Rule Engine 구조도

- 본 연구에 사용된 데이터의 가시적으로 분류할 수 있는 항목이란 성별, 연령, 교육 정도, 혼인 상태, 연간 가계 수입 이며 다음 표1과 같이 정량화 할 수 있다.

표 1. 가시적으로 분류할 수 있는 항목 및 세부 분류

항목	세부 분류
Sex	Female
"	Male
Age	20세 미만
"	20세 이상 – 30세 미만
"	30세 이상 – 40세 미만
"	40세 이상 – 50세 미만
"	50세 이상 – 60세 미만
"	60세 이상
Educational Background	6년 이하
"	7년 이상 - 9년 이하
"	10년 이상 – 12년 이하
"	13년 이상 – 16년 이하
"	17년 이상
Marital Status	Single
"	Married
"	Divorced
"	Widow or Widower
"	Co-habitation
Monthly Household Income	Under US $ 100
"	$100 – Under US$500
"	$500 – Under US$1000
"	$1000 – Under US2000
"	$2000 – Under US$3000
"	$3000 – Under US$4000
"	$4000 – Under US$5000
"	$5000 and more than US5000

- 예를 들어 최소 지지도(min. support) 0.5 이고 성별 연령 항목으로 구분하도록 지시되면 성별 2개 세부 분류와 연령 6 개 세부 분류에 대해 12개의 클러스터로 구분되어 질 수 있으며, Association Engine 에서는 각 해당 전체 레코드 수를 재 계산한 후 유동적 지지 빈도 수에 의해 빈발 항목 집합을 생성해 낸다.
- 사용되어진 연관 규칙 알고리즘은 Apriori algorithm 이며 주어진 최소 신뢰도(min. confidence)에 따라 규칙을 생성한다. 이때 생성 되어 지는 규칙은 각 클러스터 별 즉, 12개의 클러스터에 해당되는 다차원 연관 규칙(N-dim. Association rule) 이 생성되어진다.

3. 다차원 연관 규칙 마이닝

본 연구에서 제시한 다차원 연관규칙 마이닝 시스템은 가시적 분류 기준을 차원(dimension)으로 하여 일차원에서가 아닌 다차원(n-Dim.)에서의 연관 규칙을 탐사함으로써 보다 정확한 규칙을 탐사할 수 있다.

다음 그림2 는 성별, 교육 정도, 연령을 기준으로 한 3D 형태의 연관 규칙 모델의 예이며 그림 3은 실험 결과 탐사 된 규칙이다.

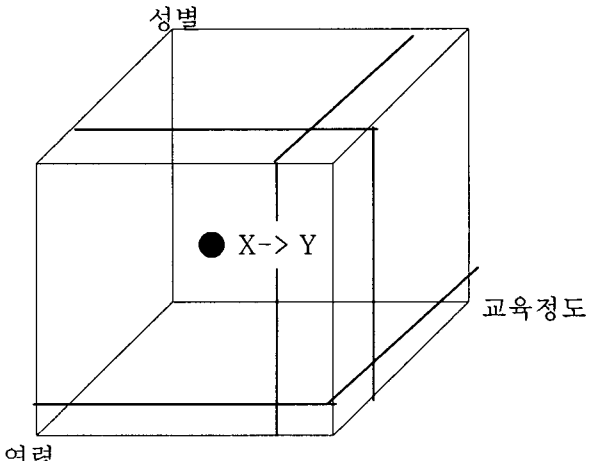

그림 2. 3D형태의 연관 규칙 모델의 예

confidence	cluster1	cluster2	cluster3	item1
1	Female	20세 미만	10 - 12 년	Do you intend to learn ICT to use it in future professional work?
1	Female	20세 미만	10 - 12 년	Do you intend to learn ICT to use it in future professional work?
1	Female	20세 미만	10 - 12 년	Do you intend to learn ICT to use it in future professional work?
0.9736842	Female	20세 미만	10 - 12 년	Marital sfatus
0.9714286	Female	20세 미만	10 - 12 년	Cell phone/ Mobile phone
0.9714286	Female	20세 미만	10 - 12 년	Cell phone/ Mobile phone
0.96666664	Female	20세 미만	10 - 12 년	TV

range1	item2	range2
Yes	I believe that knowledge and information are essential to my living.	Always
Yes	Do you use a computer?	Yes
Yes	Marital status	Single
Single	Video player	Ownership
Usage	Cell phone/ Mobile phone	Ownership
Usage	Video player	Ownership
Ownership	Marital status	Single

그림 3. 생성된 3D 연관 규칙

이와 같이 다차원 연관규칙 마이닝은 규칙을 세분화 함에 따라 일반 연관규칙을 보다 신뢰성 있는 규칙을 생성할 수 있다.

예를 들어, Do you intend to learn ICT to use it in future professional work?라는 질문에 yes 라고 응답한 것을 A라고 하고 Do you use a computer? 라는 질문에 yes 라 대답한 것을 B 라 할 때, 연관 규칙 A->B 에 대하여

- 기존 정량적 연관 규칙에서의 연구는 0.986692의 신뢰도를 보였으며
- 일차원 클러스터링 후의 연관규칙을 탐사한 결과
  - 성별로 클러스터링 한 경우 여성인 경우 신뢰도는 0.99333334
  - 연령으로 클러스터링 한 경우 20세 미만인 경우 0.9907407의 신뢰도를 40세 이상인 경우 0.9811321의 신뢰도를 나타내었다
- 이차원 클러스터링 후의 연관규칙을 탐사한 결과
  - 성별과 연령으로 클러스터링 한 경우 여성이면서 20세 미만의 경우 신뢰도는 1, 여성이며 30세 이상의 경우 신뢰도는 0.9883721, 남성이면서 30이상인 경우 0.9726027의 신뢰도를 나타내었다.
  - 연령과 교육 정도로 클러스터링 한 경우 20세 미만이면서 교육 정도가 10년-12년 사이인 경우 신뢰도는 1 이였고 40세 이상 이면서 교육 정

도가 13-16년 사인인 경우 신뢰도는 0.97959185로 나타났다.

- 성별, 연령 그리고 교육정도의 삼차원 클러스터링 후 연관규칙의 신뢰도는
  여성, 20세 미만, 10년-12년 : 1
  여성, 20세이상-30세미만, 10년-12년 : 1
  여성, 20세이상-30세미만, 6년 이하 : 1
  여성, 40세이상-50세미만, 17년-20년 : 1
  남성, 20세 미만, 6년 이하 : 1
  남성, 20세 미만, 10년-12년 : 0.9583333
  남성, 60세 이상, 13년-16년 : 1
  로 요약 되었다.

탐사된 다차원 연관규칙에서 보여지듯이, 차원을 높여갈수록 기존의 연관규칙에서 보다 강력하고 세분화된 연관규칙이 탐사되어짐을 알 수 있다.

다음 그림4 는 본 논문에서 연구한 다차원 연관규칙 마이닝을 위한 시스템의 구현 화면이다. 본 실험은 최소 지지도 0.5 최소 신뢰도 0.95로 실험 되었으며 성별, 연령, 교육정도의 각 기준에 의해 60개의 클러스터를 형성한 후 형성된 클러스터로부터 최소 지지도와 최소 신뢰도를 기준으로 다차원 연관규칙을 생성한다.

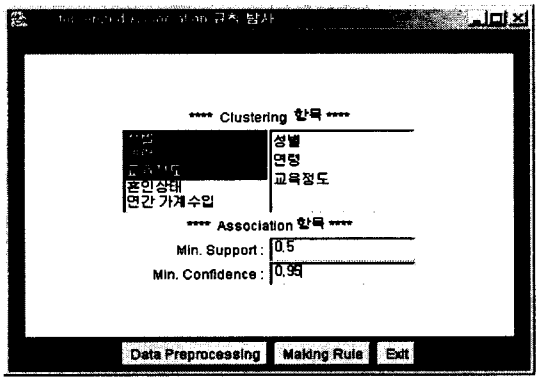

그림 4. 구현 화면 1

4. 결론

본 논문에서는 인문 사회과학 분야의 방대한 설문 데이터를 처리하기 위해 설문데이터 데이터 마이닝 시스템을 설계하고 데이터 간의 연관규칙을 찾아냄에 있어 직관적으로 분류될 수 있는 기준에 따라 클러스터링을 실행하여 데이터를 분류한 후 각 클러스터로부터 다차원 연관 규칙을 탐사하는 시스템을 제안하였다. 향후 계속되어야 할 연구로는 각 클러스터 간의 관련성 연구와 더불어 탐사 된 다차원 연관규칙에 신경망, 결정 트리, SOM 등의 다른 데이터 마이닝 알고리즘을 적용하여 데이터마이닝 모듈을 개선하는 방안들이 남겨져 있다.

5. 참고 문헌

[1] Kantardzic, M (2003). *Data Mining –Concepts, Method, And Algorithms, IEEE Computer Society,* Wesley Interscience

[2] Usama M. Fayyad, Gregory Piatetsky-Shapiro and Ramasamy Uthurusamy, (1996). "Advanced in Knowledge Discovery and Data Mining" *AAAI Press / The MIT Press pp.307-328*

[3] *Joseph P. Bigus and Jenifer Bigus,(2001) Constructing Intelligent Agent Using Java second Eedition, Wiley Computer Publishing,*