전자 상거래 에이전트를 위한 연관 규칙 발견 및 확장

문 홍 기, 이 수 원
숭실대학교 컴퓨터학부

Association Rule Discovery & Expansion for Electronic Commerce Agents

Hong Gi Moon, Soowon Lee
School of Computing, SoongSil University

요 약

대용량 데이터베이스의 데이터로부터 지식을 발견하는 방법으로 사용되고 있는 연관 규칙 발견은 기존에는 알려지지 않았던 지식을 찾아 이를 이용할 수 있는 형태로 제공된다. 하지만, 제공되는 형태는 단순한 데이터베이스에 포함되어 있는 정보만을 이용하여 보여주므로, 특정한 부분에만 제한적으로 활용된다. 따라서, 본 연구에서는 데이터로부터 연관 규칙을 발견하여 이를 개념 계층구조를 이용하여 일반적인 규칙으로 확장하는 방법을 제안한다. 또한, 발견된 규칙을 기반으로 전자 상거래 에이전트를 위해 어떻게 활용될 수 있는지를 제안한다.

1. 서론

최근 데이터마이닝은 대용량의 데이터베이스를 가지고 있는 기업에 대해 효율적인 마케팅 전략을 세울 수 있는 정보를 제공할 수 있기 때문에 주목을 받고 있다. 이는 데이터베이스 분야에서는 데이터마이닝이라 부르지만, 인공지능 분야에서는 데이터에 대한 사용 가능한 지식을 추출한다는 의미로 지식 발견(Knowledge Discovery in Database)이라 부르고 있다. 연관 규칙 발견은 데이터베이스로부터 찾기 힘들고, 암시적인 규칙을 찾는 것으로 현재 크로스 마케팅, 첨부 메일, 카탈로그 디자인, 구매자의 호응도 조사, 구매자 분석 등에서 활용되는 지식으로의 역할을 하게 된다. 하지만, 발견된 규칙들은 단지 아이템들의 정보에 기반하기 때문에 이 아이템들의 의미적인 정보는 무시하게 되므로, 규칙들 간의 관계 등이 결여된 형태의 지식으로 구축된다. 개념들의 계층 구조를 이용하여 규칙에 포함된 아이템들의 관계를 유추하여 이를 규칙 확장에 이용한다면 보다 설득력 있는 지식으로의 역할을 하게된다.

전자상거래 에이전트는 일반적으로 (1) 구매자의 상품 탐색 과정을 도와주고 추천해 주는 상품 추천 에이전트(recommendation agent), (2) 구매자의 취향 및 특성에 따라 상품에 대한 카탈로그 정보를 푸쉬(push)하여 제공하는 통지 에이전트(notification agent), (3) 구매자가 지정한 특정 상품에 대하여 여러 판매 사이트의 가격 및 조건 등을 비교하여 제시하는 비교 쇼핑 에이전트(comparison shopping agent), (4) 구매자와 판매자가 제시한 여러 조건을 만족시킬 수 있도록 협상하고 가장 적합한 구매자와 판매자를 중개해 주는 협상 에이전트(negotiation agent)등으로 구분된다. 본 연구에서는 연관 규칙 발견 알고리즘 중에 최신 알고리즘인 DHP(Direct Hashing and Pruning)[1]을 사용하여 연관 규칙 발견 모듈을 구현하였고, 발견된 규칙들에 대해 개념정보를 가지고 있는 WordNet의 정보를 이용하여 규칙들을 일반화하여 전자 상거래 에이전트를 위한 지식베이스로의 역할을 하게된다. 연관 규칙 발견 모듈은 Java 2 SDK로 구현하였고, 연관 규칙은 CLIPS 규칙 형태로 생성되어 추후에 CLIPS 추론 엔진에서 사용할 수 있도록 구성되었다.

2. KDD (Knowledge Discovery in Database)

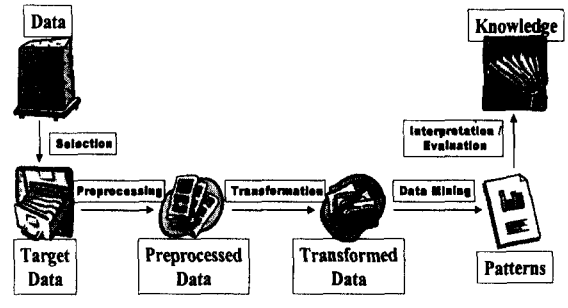

그림 1 : 지식 발견 프로세스

지식 발견 프로세스는 그림 1과 같이 데이터에 대해 필요한 부분만을 선택하여 이를 전처리 과정을 통해 지식을 추출하기 쉬운 형태로 만들고, 다시 필요에 따라 자료 형태를 바꾸어 주게 된다. 패턴을 쉽게 추출할 수 있는 구조로 변환된 자료에 대해 데이터마이닝 알고리즘을 사용하여 패턴을 찾고, 이를 해석하고 평가하여 지식으로 활용한다. 기존의 지식 발견에 대한 활용은 단순히 전문가의 전략 수립에 필요한 정보로 활용되기 때문에 발견된 지식은 수동적인 성격을 가지게 되어 전문가가 어떤 결정을 내리기 전에는 쓸모 없는 정보가 된다. 따라서, 본 연구에서는 발견된 지식을 지식베이스의 규칙의 형태로 저장하여 새로운 규칙이 발견될 때마다 규칙의 형태로 지식이 구축되고, 구축된 지식을 전자 상거래 에이전트가 활용할 수 있는 방안을 제시한다.

3. 연관 규칙

연관 규칙 발견 알고리즘으로 Apriori, OCD, SETM, DHP 알고리즘이 연구되었고, 성능 면에서 해시테이블을 사용하는 DHP 알고리즘이 가장 우수하다[2][3][4][5]. 따라서, 본 연구에서는 DHP 알고리즘을 사용하여 연관성을 추출한다.

T를 현재까지 발생한 모든 트랜잭션이라 하고, I를 m개의 상품들 또는 구매자들이라 하면 T와 I는 다음과 같이 표현된다.

$$T = \{t_1,\ldots,t_n\}$$

$$I = \{i_1,i_2,\ldots,i_m\}$$

어떤 트랜잭션 t에 상품 또는 구매자 w가 포함되어 있을 때, t(w)는 다음과 같이 표현된다.

$$t(w) = 1$$

w는 트랜잭션 t에 존재하는 상품 또는 구매자 중에 하나이다. 또한, W가 상품들의 집합이라고 했을 때, t(W)는 트랜잭션 t에 대해 W에 있는 모든 상품 또는 구매자가 존재함을 말한다.

$$t(W) = 1 : t(w) = 1, \text{ 모든 } w \in W$$

W : I에 속하는 상품들 또는 구매자들의 부분집합

$$(X) = \{i \mid t_i(X)=1\}$$

(X) 는 X가 포함하는 모든 상품 또는 구매자들을 포함하고 있는 모든 트랜잭션의 집합이 된다. 여기서, X : |(X)|≥σ 이면 X를 σ-covering이라 부른다. 따라서, 모든 트랜잭션에 대해 다음과 같은 법칙을 발견할 수 있다.

W ⇒ B : T에 대해 W⊆R 이고 B⊆R\W를 만족하는 연관 규칙

W : 연관 규칙의 왼쪽 집합

B : 연관 규칙의 오른쪽 집합

W ⇒ B는 지지도(support threshold)와 신뢰도(confidence threshold)에 의해 발견된다. 지지도는 이 법칙이 적용되는 트랜잭션의 수를 나타내고, 신뢰도는 연관 규칙의 왼쪽 집합이 가리키는 전체 트랜잭션 개수 중에서 이 법칙을 만족하고 있는 트랜잭션의 개수를 나타낸다. 따라서, 지지도와 신뢰도에 의해 다음의 연관 규칙이 발견된다.

T 는 W⇒B 을 만족한다.

단, $W \cup B$: σ-covering (i.e., $|(W \cup B)| \geq \sigma$) σ : 지지도

$|(W \cup B)|/|(W)| \geq \gamma$ γ : 신뢰도

연관 규칙 발견은 간단히 2가지 알고리즘에 의해 쉽게 발견될 수 있다. 하지만, 이들 알고리즘은 모든 아이템에 대해 가능한 부분집합들을 구해야 하므로 NP-hard로 취급된다. 따라서, 성능을 개선하기 위해 병렬 알고리즘과 같은 다양한 방법에 대한 연구들이 활발히 진행되고 있다.

첫 번째 알고리즘은 왼쪽 집합의 모든 가능한 부분집합을 생성하는 알고리즘으로 트랜잭션 σ개 이상에서 발생하는 아이템들의 가능한 모든 집합을 찾아 이를 후보 아이템 집합들로 등록시키는 σ-cover 알고리즘이다. 표 1은 σ-cover 알고리즘을 나타낸다.

σ-cover algorithm에 의해 생성된 집합은 연관 규칙의 왼쪽 집합에 관한 제약 사항을 만족하는 집합 Ls 이다. 두 번째 알고리즘은 바로 이 집합에 대해 연관성을 가진 오른쪽 집합을 찾는 알고리즘으로 연관성이 있는 모든 아이템들의 부분집합을 찾게된다. 이 집합들 간의 관계가 바로 연관 규칙이 된다. 구체적인 알고리즘은 표 2와 같다.

```
Candi = { {A} | (A) ≥ σ }
i = 1
While Candi ≠ ∅ do
        Candi+1 = { S1∪S2 | S1, S2  Candi, | S1∪S2 | = i + 1,
            All subsets of S1∪S2 are in Candi }
end do
Evaluate  ⋃_i Candi
```

표 1 : σ-cover 알고리즘

```
For all D∈Ds
  For all X∈Ls do
    if X⊆K(D) then
      B = 현재 연관 규칙 오른쪽 집합을 만족하는 X에
          포함된 상품 또는 구매자
      X와 B의 모든 부분 집합들에 대해 동시 발생
          카운터를 증가시킴
    end if
  end do
end do
```

표 2 : 연관 규칙 발견 알고리즘

4. 연관 규칙 발견 모듈

본 연구에서 구현된 연관 규칙 발견 모듈은 구매 트랜잭션에 포함된 아이템들의 속성에 따라 연관 규칙을 찾아내는 모듈로 전자 상거래 에이전트가 사용할 수 있는 지식을 추출하여 추론 가능한 CLIPS rule로 표현하여 지식베이스에 저장된다.

가상의 트랜잭션이 표 3, 표 4 와 같은 테이블 형태로 데이터베이스에 저장되어 있다고 할 때, 연관 규칙 모듈은 그림 2와 같은 결과를 보여주게 된다.

ID	goods
1	GUESS_jean1990B, GUESS_shirt193Y
2	CK_jean130, NAUTICA_dshirt130
3	LEVIS_jean505, POLO_cap10w
4	QUICKSILVER_pants120
5	ELCANTO_g350B, ELCANTO_p66B
6	LEE_jean345, BEANPOLE_shirt14g
7	SWATCH_s735, NIKE_shirt319
8	ADIDAS_tshoes119w, ADIDAS_shirt308w
9	PROSPECS_shoes201b, NIKE_shirt319
10	POLO_dshirt336r, NAUTICA_dshirt131
11	HUNT_pants109w, LEVIS_jean505
12	BODYGUARD_pantie203, GUESS_cap660
13	DAEWOO_solo600
14	LG_tv340, LG_ahafree120
15	SAMSUNG_tv2000, SAMSUNG_vcr3000
16	LG_cam120, LG_mymy210

표 3 : 트랜잭션 테이블

ID	kind
GUESS_jean1990B	jean
GUESS_shirt193Y	shirt
GUESS_cap660	cap
CK_jean130	jean
LEVIS_jean505	jean
NAUTICA_dshirt130	dress shirt
NAUTICA_dshirt131	dress shirt
QUICKSILVER_pants120	pants
ELCANTO_g350B	shoes
ELCANTO_p66B	purse
LEE_jean345	jean
POLO_dshirt336r	dress shirt
POLO_cap10w	cap
BEANPOLE_shirt14g	shirt
SWATCH_s735	watch
NIKE_shirt319	shirt
ADIDAS_tshoes119w	running shoes
ADIDAS_shirt308w	shirt
PROSPECS_shoes201b	running shoes
HUNT_pants109w	pants
BODYGUARD_pantie203	pantie
DAEWOO_solo600	notebook
LG_tv340	television
LG_ahafree120	cassette
SAMSUNG_tv2000	television
SAMSUNG_vcr3000	vcr
LG_cam120	camcorder
LG_mymy210	cassette

표 4 : 상품 테이블

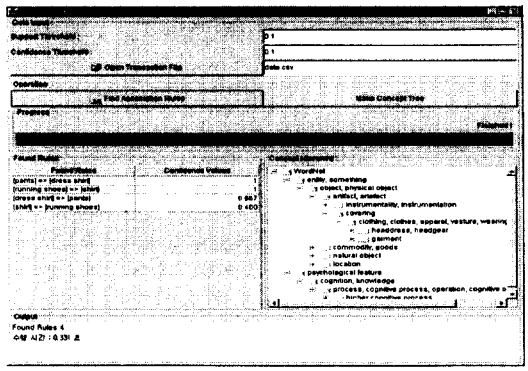

그림 2 : 연관 규칙 발견 모듈

5. 연관 규칙 확장

연관 규칙 발견 모듈에서 추출되는 연관 규칙들은 아이템들의 종류와 같은 한 레벨 위의 상위 개념 정도가 포함되어 있을 뿐이다. 따라서, 발견되는 지식들은 아이템에 의존적일 뿐 아니라 특정부분에만 한정되어 사용될 수밖에 없는 연관 규칙이 된다. 본 연구에서는 이러한 결점을 없애기 위해 프린스톤 대학에서 개념을 중심으로 계층적으로 의미망을 구성한 WordNet을 이용하여 발견된 규칙들에 대해 일반화 작업을 통하여 규칙을 확장하여 구성하도록 하였다[6][7].

현재 존재하고 있는 아이템들에 대해 이 아이템의 종류를 포함하고 있는 상위 개념들을 WordNet을 이용하여 개념 계층 구조로 구성하게 된다. 표 5는 연관 규칙 모듈에 의해 발견된 규칙들을 나타낸다.

Found Rules	Confidence
[running shoes] => [shirt]	1
[shirt] => [jean]	0.400
[shirt] => [running shoes]	0.400
[jean] => [shirt]	0.400

표 5 : 발견된 연관 규칙

발견된 규칙들에 대하여 확장을 하기 위해서 WordNet에서 제공하는 정보를 이용하여 running shoes, shirt, jean 에 대한 개념들 간의 관계를 그림 4와 같이 구성하게 된다.

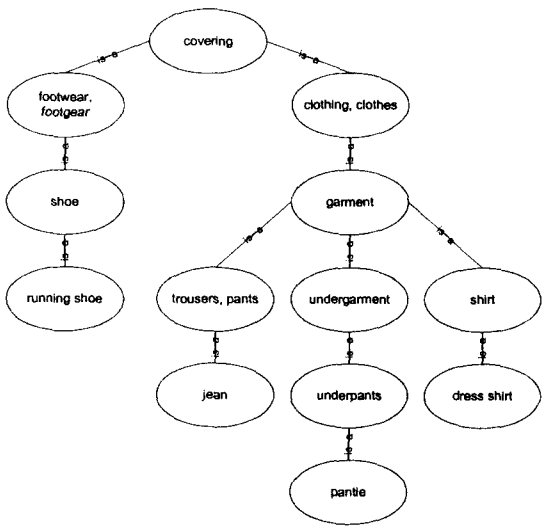

그림 3 : WordNet에서 추출한 개념 계층 정보

따라서, 발견된 각각의 규칙들에 대해 개념들의 관계를 이용하여 일반적인 개념들을 사용하여 규칙의 신뢰도를 높이거나 새로운 규칙을 찾게 된다. 예를 들면, dress shirt는 shirt와 is-a 관계에 있으므로 모든 dress shirt는 shirt라고 봐도 무관하다. 따라서, dress shirt를 shirt로 바꾸어 연관성을 찾으면 표 6과 같이 규칙의 신뢰도를 높이게 된다.

Found Rules	Confidence
[running shoes] => [shirt]	1
[jean] => [shirt]	0.600
[shirt] => [jean]	0.375
[shirt] => [running shoes]	0.25

표 6 : 확장된 연관 규칙 (1)

또한, jean도 pants와 is-a 관계에 있기 때문에 jean을 pants로 바꾸고 연관 규칙을 발견하면 표 7과 같은 새로운 규칙을 찾을 수 있다.

Found Rules	Confidence
[running shoes] => [shirt]	1
[pants] => [shirt]	0.429
[shirt] => [pants]	0.375
[shirt] => [running shoes]	0.25

표 7 : 확장된 연관 규칙 (2)

위와 같이 연관 규칙을 찾고, 규칙을 확장하는 궁극적인 목적은 전자 상거래 에이전트의 중요한 지식이 되어 능동적인 마케팅에 참여할 수 있도록 개발하기 위해서이다. 현재 구현된 연관 규칙 발견 모듈은 순수 100% 자바로 구현되어 에이전트가 필요하다면 언제나 사용할 수 있고, 자바 기반의 다중 에이전트 시스템에서 그 활용 가치는 더 높아질 것으로 보인다.

6. 결론 및 향후 과제

본 논문에서는 전자 쇼핑몰에서 일어나는 구매 행위에 대해 연관 규칙 발견 모듈을 구현하였고, 발견된 연관 규칙의 확장을 위해 WordNet을 사용하여 규칙을 일반화하여 이를 에이전트가 활용할 수 있도록 구성하였다. 현재, 데이터베이스하의 지식 발견에 대해서는 많은 방법들이 연구되었고, 실용화된 시스템도 상당수 나와있다. 하지만, 이러한 시스템들은 모두 지식 발견 수행의 결과만을 사람에게 보여줄 수 있는 형태로 제공되었기 때문에 의사 결정에 도움이 될 뿐, 결과에 의거한 적합한 행위를 하기에는 어려움이 따를 수 있다. 따라서, 결과에 능동적인 행위를 할 수 있는 개체로 전자 상거래 에이전트가 사용하게 되었고, 지식 추출에 의한 결과를 에이전트가 바로 이용할 수 있게 하는 방법으로 연관 규칙을 사용한 지식 베이스를 구축하였다.

데이터베이스 내부에 존재하고 있는 지식 추출은 단지 연관 규칙 발견에만 있는 것은 아니므로 다른 방법에 대해서도 에이전트의 지식으로서 활용될 수 있는 방안을 연구하고 이를 적용시켜야 할 것이다. 또한, 전자 상거래 뿐 아니라 다른 분야에서도 이를 활용할 수 있는 곳에 적용하여 생산성을 높이는 연구도 병행되어야 한다.

참고 문헌

[1] J. S. Park, M.-S. Chen and P. S. Yu. An Effective Hash Based Algorithm for Mining Association Rules. *Proceedings of ACM-SIGMOD Conference on Management of Data*, pages 175-186, San Jose, California, May 1995.

[2] J. L. Han and A. W. Plank. Background for Association Rules and Cost Estimate of Selected Mining Algorithms. *Proceedings of ACM-CIKM*, pages 73-80, Rockville, Maryland, August, 1996

[3] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *In Proc. of the ACM SIGMOD Conference on Management of Data*, pages 207-216, Washington, D.C., May 1993.

[4] H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient Algorithms for Discovering Association Rules. In *KDD-94: AAAI Workshop on Knowledge Discovery in Databases*, pages 181-192, Seattle, Washington, July 1994

[5] R. Feldman and H. Hirsh. Finding Association in Collections of Text. In *Machine Learning and Data Mining*. John Wiley & Sons Ltd, 1997

[6] Ramakrishnan Srikant and Rakesh Agrawal. Mining Generalized Association Rules. *Proceedings of 21st VLDB Conference* Zurich, Swizerland, 1995

[7] E. M. Voorhees. Using WordNet to Disambiguate Word Senses for Text Retrieval. *Proceedings of ACM-SIGIR*, pages 171-180, Pittsburgh, Pannsylvania, June, 1993