Journal of the Korea Industrial Information Systems Research Vol. 20 No.5, Oct. 2015 :61-69
http://dx.doi.org./10.9723/jksiis.2015.20.5.061 ISSN:1229-3741

효율적인 연관규칙 감축을 위한 WT-알고리즘에 관한 연구

(A Study on WT-Algorithm for Effective Reduction of Association Rules)

박 진 희[1], 피 수 영[2]*
(Jin-Hee Park and Su-Young Pi)

요 약 매일 각종 모바일 디바이스와 온라인, 소셜네트워크서비스 등에서 쏟아지는 데이터로 인해 정보의 홍수를 넘어 과부하 상태에 있다. 이미 생성되어 있는 기존 정보들도 있지만 시시각각 새롭게 생겨나고 있는 정보들이 헤아릴 수 없을 정도이다. 연관분석은 이러한 정보들 속에서 나타나는 항목의 발생 빈도수가 최소 지지도보다 큰 빈발항목집합(Frequent Item set)을 찾는 방법이다. 항목의 수가 많아짐에 따라 규칙의 수도 기하급수적으로 늘어나므로 원하는 정보를 찾기가 어려운 단점이 있다. 따라서 본 논문에서는 트랜잭션데이터 집합을 Boolean 변수 아이템으로 나타내었다. 논리함수를 간소화하는데 사용되는 Quine-McKluskey의 방법으로 알고리즘화하여 각 항목에 가중치를 부여한 WT-알고리즘을 제안한다. 제안한 알고리즘은 항목의 개수와 관계없이 간략화가 가능한 장점으로 인하여 불필요한 규칙을 감소시켜 데이터마이닝 효율을 향상시킬 수 있다.

핵심주제어 : 빅데이터, 연관규칙, 가중치, 데이터마이닝

Abstract We are in overload status of information not just in a flood of information due to the data pouring from various kinds of mobile devices, online and Social Network Service(SNS) every day. While there are many existing information already created, lots of new information has been created from moment to moment. Linkage analysis has the shortcoming in that it is difficult to find the information we want since the number of rules increases geometrically as the number of item increases with the method of finding out frequent item set where the frequency of item is bigger than minimum support in this information. In this regard, this thesis proposes WT-algorithm that represents the transaction data set as Boolean variable item and grants weight to each item by making algorithm with Quine-McKluskey used to simplify the logical function. The proposed algorithm can improve efficiency of data mining by reducing the unnecessary rules due to the advantage of simplification regardless of number of items.

Key Words : Big data, Association Rules, Weight, Data mining

1. 서 론

데이터는 거대한 정보의 패러다임이다. 빅데이터는 크게 보면 우리의 생활로부터 생산되는 모든 정보를 비롯하여 작은 행동이나 습관에서 비

* Corresponding Author : sypi@cu.ac.kr
Manuscript received March 06, 2015 / revised July 31, 2015 / accepted October 07, 2015
1) 대구한의대학교 교양학부, 제1저자
2) 대구가톨릭대학교 교양교육원, 교신저자

롯되는 모든 정보를 포함한다. 매일 각종 모바일 디바이스와 온라인, 소셜네트워크 서비스(Social Network Service) 등에서 쏟아지는 데이터로 인해 정보의 홍수를 넘어 과부하 상태에 있다. 이미 생성되어 있는 기존의 정보들도 있지만 시시각각 새롭게 생겨나고 있는 정보들이 헤아릴 수 없을 정도이다. 빅데이터가 저장되는 공간은 컴퓨터의 서버이며 서버로부터 공유되는 정보는 결국 사람에 의해 생산, 확산, 가공된다. 사람이 정보를 수집하고 소유하고 공유한다. 정보화 시대의 자본은 곧 정보라고 했다. 그런데 이제는 정보를 수집하고 공유하는 일련의 과정 속에서 새로운 행동 양식이 나타나기 시작했고, 관련 서비스들도 새롭게 시장을 선점하고 있다.

빅데이터 기술의 발전은 다변화된 현대사회를 보다 정확하게 예측하고 효율적으로 작동하도록 정보를 제공하는 동시에 과거에는 불가능했던 기술을 가능케 하였다. 이러한 빅데이터 분석기법은 국가차원에서의 사회, 경제, 정치, 문화, 과학기술 등 여러 분야에 활용될 수 있다. 빅데이터 분석기술은 오프니언 마이닝, 감정분석, 데이터마이닝, 딥러닝 등이 있다. 빅데이터 분석을 위해서는 먼저 데이터마이닝 기술로 방대한 양의 데이터 속에서 가치 있는 정보를 찾는 것이 선행되어야 하는데, 빅데이터와 관련된 데이터마이닝 기법으로는 텍스트 마이닝, 평판 분석, 군집 분석, 연관성 분석 등이 있다. 이러한 데이터마이닝의 기법 중 가장 많이 사용되고 있는 것이 연관성 분석이다. 연관성 분석은 데이터베이스 내에 존재하는 항목들 간의 관계를 기술하는 것으로 규칙의 표현이 간략하며 사용자에게 데이터에 관한 유용한 정보를 줄 수 있다. 그렇지만 이러한 연관성 분석은 항목의 수가 늘어날수록 규칙의 수도 기하급수적으로 늘어나게 되므로 고속처리 방식의 실현이 중요한 과제가 되고 있다[1-3].

따라서 본 논문에서는 연관규칙 마이닝에서 방대한 데이터베이스에 저장되어 있는 데이터로부터 유용한 정보 및 지식을 추출하는데 가장 잘 알려져 있고, 많은 연구가 이루어지고 있는 연관규칙을 기반으로 하였으며 Boole 대수 간소화로 알려진 Quine-McKluskey 방법[4]을 알고리즘화하여 연관규칙 마이닝에 기반한 규칙 감축방법을 제안하였다. 감축방법은 변수의 개수와 관계없이 간략화가 가능한 장점이 있다. 연관규칙을 감축하는 알고리즘의 첫 번째 단계는 전처리 단계로서 데이터를 분석에 적합한 형태로 변환하고 중복된 속성들을 제거하는 단계로서 각 항목의 속성에 따라 가중치를 부여하여 항목의 가중치가 낮은 규칙은 제거하는 방법으로 규칙을 감축하였다. 두 번째 단계는 T-알고리즘[5]을 이용하여 변수의 개수에 관계없이 연관규칙을 감축시켜 준다.

제안된 알고리즘의 타당성을 검증하기 위해 기존 연관규칙 탐색방법인 Apriori 알고리즘과 이전에 제안한 T-알고리즘과 비교하였을 때 제안한 알고리즘의 효율성이 높음을 실험을 통하여 확인한다.

2. 관련연구

2.1 Apriori 알고리즘

데이터에 숨겨진 패턴을 탐색하는 데이터마이닝에서 가장 많은 연구가 이루어진 분야가 연관성 분석이다. 이것은 대용량의 데이터베이스에서 어떤 사건들이 함께 발생하거나 또는 하나의 사건이 다른 사건을 암시하는 것과 같은 사건 간의 상호관계를 연관시킨 것으로 항목 X와 Y사이의 X→Y 형태의 규칙을 찾아낸다. 여기서 항목은 일반적으로 이진속성을 가지며 항목 X가 나타나면 항목 Y도 나타날 가능성이 높다는 연관관계를 나타낸다. 이와 같이 연관규칙에서 사용되는 입력데이터는 주로 이진형식을 가지는 경우가 많다[6].

연관규칙 마이닝에서는 다량의 트랜잭션 데이터를 취급하기 때문에 고속처리방식의 실현이 중요하므로 Apriori 알고리즘이 많이 사용되고 있다. Apriori 알고리즘은 구현이 간단하고 성능 또한 만족할 만한 수준을 보여주는 알고리즘으로써 강한 연관성을 갖는 항목들을 발견하고, 각 단계 빈발 항목들을 발견하는데 초점을 둔다. 각 단계에 빈발 항목들의 후보 항목집합을 구성하고 난 후 각 후보 항목집합의 발생빈도수를 구하고 사

용자가 정의한 최소 지지도의 최소 신뢰도를 기초로 하여 빈발 항목집합들을 정의한다. Apriori 알고리즘은 많은 항목들과 트랜잭션들이 알고리즘의 후반부 패스들에서 더 이상 필요 없음에도 불구하고 항상 각 단계마다 전체 데이터 셋을 검색해야 한다. 비빈발 항목이나 트랜잭션들을 제거하는 것은 후보가 될 가능성이 없는 집합들을 카운트하기 때문에 빈발항목 측정에 시간과 비용이 많이 드는 단점이 있다[7-8].

연관규칙의 지지도(Support) S와 신뢰도(Confidence) C를 식(1)과 (2)와 같이 정의한다. 여기서 N은 데이터베이스의 모든 트랜잭션의 수, T는 트랜잭션을 나타내고 P는 확률을 나타낸다.

$$S = \frac{T[X \cup Y]}{N} = P(X \wedge Y) \quad (1)$$

$$C = \frac{T(X \cup Y)}{T(X)} = P(Y \mid X) \quad (2)$$

지지도 S는 식(1)과 같이 데이터베이스 중에 X와 Y가 동시에 출현하는 확률이다. 신뢰도 C는 식(2)와 같이 X에서 Y가 발생할 조건부 확률이다. X→Y라는 연관규칙은 조건부 X가 성립했을 때 후건부 Y가 성립할 확률이 신뢰도로 주어진다. 지지도는 데이터베이스 중의 어느 정도의 트랜잭션에 대해 이 규칙을 적용할 수 있는지의 비율을 나타내고 있고 규칙의 범용성의 지표가 된다[9-10].

2.2 Quine-McKluskey 방법

불대수(Boolean Algebra)의 간략화를 위한 Karnaugh Map방법은 구조적인 특성상 변수의 수가 적은 경우에 적합하다. 4개 이상의 변수들에 대하여 Karnaugh Map를 사용할 경우 계산이 매우 복잡해지므로 최적화 된 Boolean 함수를 구하기가 어렵게 된다[11]. Karnaugh Map 보다 더 간단하고 체계적인 방법으로 Boolean 함수의 간략화를 위한 방법이 Quine-McKluskey 방법이며 도표에 의해서 간략화를 행하므로 Tabular 방법이라고도 한다. 이 방법은 변수의 개수와 관계없이 간략화가 가능한 장점이 있으므로 Boolean 식을 자동으로 간략화 시키는데 적합하다. Quine-McKluskey 방법의 간략화방법은 최적의 곱의 합의 형태를 얻기 위하여 최소 항의 결합에 의해서 리터럴 수를 줄여나가는 과정을 반복하면 되는데 이 과정은 두 단계로 나눌 수 있다. 첫 번째 단계는 XY+XY′=X의 정리를 적용하여 각 항에 있는 리터럴의 수를 최대로 줄여 결과의 항인 주항(prime implicant)을 얻는다. 주항을 찾기 위해서 각각의 최소 항들은 오직 한 변수만 다른 항들과 비교된다. 즉, 임의의 두 최소항의 비교에서 한 변수만이 다르다면 서로 다른 변수들은 제거되고 나머지 변수로 구성된 새로운 적의 항을 얻는다. 이러한 방법을 이용해서 최소 항에 대해서 비교를 행하여 새로운 논리곱의 항을 얻는데 이러한 비교과정은 새롭게 구해진 모든 항에 대해서도 반복한다. 위의 과정을 반복한 후에 조합되지 않은 모든 항과 남아있는 항이 주항이 된다. 비교과정 중에 요구되는 비교의 횟수를 줄이기 위해서 1의 개수가 같은 최소 항 끼리 그룹화를 행하고 1의 개수 차이가 1인 모든 그룹에 대해서 각 최소 항 끼리 비교한다.

두 번째 단계는 주항 표(prime implicant table)를 사용하여 최소화된 함수를 형성하고 있는 주항을 선택하고 선택된 항들을 논리합하여 최소의 리터럴 수를 갖는 간략화 된 함수를 얻는다[4].

3. WT-알고리즘을 적용한 규칙감축

3.1 연관규칙의 감축알고리즘

지식의 표현에 논리 벡터를 사용하거나 사례집합으로부터 명제를 획득하는 방법이 기계학습 분야에서 많이 연구되고 있으며 이산값으로서 표현할 수 있는 데이터를 대상으로 하고 있다. 본 논문에서는 트랜잭션 데이터집합을 Boolean 아이템으로 나타내어 WT-알고리즘(Weighted Tabular-Algorithm)을 적용하여 연관규칙을 감축하고 트랜잭션 데이터베이스 항목에 가중치를 부여하여 데이터마이닝 효율을 향상시키는 방법을 제안한

다. 본 논문에서 제안하는 연관규칙 감축알고리즘은 Boolean 함수를 간소화하는 Quine-McKluskey 간소화 방법에 가중치를 결합하여 데이터마이닝에 적용시키기 위해 알고리즘화하고 Fig. 1과 같이 나타내어 연관규칙 감축에 적용시켰다. 입력된 트랜잭션데이터는 전처리단계를 거쳐 추출된 규칙을 T-알고리즘을 사용하여 규칙을 감축한다[5]. 전처리단계에서는 트랜잭션 데이터베이스 내에 존재하는 규칙에 가중치를 주어 최소 가중치를 만족하는 규칙만을 추출하여 T-알고리즘에 적용하게 된다. 전처리단계에서는 데이터를 분석에 적합한 형태로 변환하고 중복된 데이터와 불필요한 데이터들을 제거하는 단계이다.

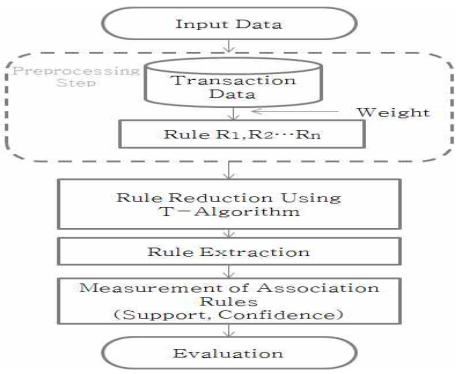

Fig. 1 Reduction Algorithm of Association Rules

단계 1에서 단계 3은 각 규칙마다 가중치를 부여하여 불필요한 규칙을 제거하는 단계로서 다음과 같은 3단계의 과정을 수행한다.

[단계 1] 전건부의 규칙 중 1의 개수가 1개인 것부터 n개인 것을 차례로 모아 정렬한다.

[단계 2] 트랜잭션 데이터베이스의 규칙 중 후건부의 값이 1인 전건부의 규칙을 찾아낸다.

[단계 3] 전 단계에서 찾아진 규칙의 가중치를 계산하여 가중치가 낮은 규칙을 제거한다. 전처리단계를 거친 규칙들을 T-알고리즘을 이용하여 감축을 수행한다. 단계 4에서 1의 개수가 n개인 항과 n+1의 항을 비교하여 1의 차이가 있는 최소 항을 골라내는 방식으로 알고리즘을 진행하게 된다. 단계 4에서 단계 7은 전처리단계를 거친 데이터들을 T-알고리즘을 적용하여 감축을 진행한다.

[단계 4] 전처리를 거쳐 정렬된 데이터끼리 비교하여 감축을 수행한다.

XY+XY′=X의 정리를 적용하여 최소 항과 비교해서 비교하는 항들 간에 서로 다른 값 최소 항들을 골라낸다. 이때 결합한 값은 don't care bit(-)와 care bit(1,0)으로 표시된다. 예를 들어, 1의 개수가 1개인 x_1이 0010이고 1의 개수가 두 개인 x_2가 1010일 때 {X_1,X_2}=-010으로 표시하여 1차 감축을 수행한다.

[단계 5] 단계 4의 작업을 1의 개수가 n개인 항과 n+1인 항을 비교하여 감축한다.

[단계 6] 1차 감축 결과를 모아서 단계 4와 단계 5의 동일한 방법으로 2차 감축을 시행한다. 단, 1차 감축에서 더 이상의 감축이 진행되지 않을 경우 2차, 3차 감축은 진행하지 않는다.

[단계 7] 최종단계로 중복되는 규칙을 제거한다.

이렇게 감축된 n개의 규칙은 규칙의 연관성을 측정하기 위하여 지지도와 신뢰도를 측정하게 된다. 본 논문에서는 총 7단계 감축단계를 거쳐 연관규칙 알고리즘을 진행하며 감축순서는 단계 1

k=1
$F_k = \{ i \mid i \in I \wedge \sigma(i) \geq N, i \notin 0 \}$ // All 1-Items Set
$S_k = \{ i \mid i \in I \wedge \sigma(i) \geq N, N > 0, i = descending \}$
repeat
for each reduction k=k+1
$S_k = T_{algrithm} - cre(F_{k-1})$ // Generation of descending Items Set
for each transaction $t \in T$ do
$S_k = \min(S_k, t)$ // All Items belonging to t(Minterm)
for each candidate item set $s \in S_k$ do
if each S_i bit compare each S_k bit then
equal bit then S_i 1 or 0
not equal bit then S_i= -
// don't care bit (-), care bit (1,0)
end for
end for
end for
$F_k = \{ s \mid s \in S_k \wedge \sigma(s) \geq N, N \neq 0 \}$
// k-reduction Items Set Extraction
$F_k = \{ s \mid s \in S_k, N \neq 0 \}$
// k-reduction Items deduplication Measurement
until $F_k = 0$
Result $= \cup F_k$

Fig. 2 Structure of T- Algorithm

에서 단계 7까지이다. 생성된 n개의 연관규칙 집합들에 대하여 지지도와 신뢰도를 계산하여 최종 평가하게 된다. T-알고리즘 감축과정을 Fig. 2에 나타내었다.

3.2 연관규칙의 가중치 부여

단계 1에서 단계 2를 거친 규칙을 베이즈이론을 이용하여 아래 식(3), (4), (5)을 이용하여 각 규칙에 대한 가중치적용을 위해 x_1에서 x_4까지 1과 0에 대한 사전확률 값을 계산한다.

$$P\ (true|x_i)+P\ (false|x_i)=1 \tag{3}$$

$$P\ (true|x_i)=\frac{n(true)}{N} \tag{4}$$

$$P\ (false|x_i)=\frac{n(false)}{N} \tag{5}$$

위의 식에서 N은 전체 트랜잭션 규칙의 수이고 $n(true)$는 true의 개수, $n(false)$은 false의 개수를 나타낸다. 각 트랜잭션 데이터규칙에 대한 사전확률 값을 이용하여 사후확률 즉, 가중치(W_p)를 식(6)을 이용하여 계산한다.

$$W_p = \frac{\sum_{i=1}^{n} x_i}{\sum_{j=1}^{n} I_j} \tag{6}$$

여기서 x는 항목을 의미하고, i는 전체 아이템의 항목수를 의미한다. 사후확률 값 즉, 가중치를 구한 후 최소 가중치를 만족하는 규칙들만을 대상으로 T-알고리즘을 이용하여 규칙을 감축한다.

3.3 규칙의 연관성 측정

연관규칙이란 하나의 거래나 사건에 포함되어 있는 항목들의 경향을 파악해서 상호연관성을 발견하는 것이다. 데이터베이스 내에서 연관규칙을 찾아내는 것은 데이터마이닝에서는 매우 중요한 일이다. 트랜잭션집합이 주어질 때, 각각의 거래는 항목의 집합이다[12-13]. 연관규칙은 형식 X→Y의 함축적인 식으로 여기서 X와 Y는 서로 상이한(disjoint) 항목집합, 즉 $X \cap Y = \varnothing$ 이다. T-알고리즘에 의해 생성된 규칙들 간의 연관성을 측정하는 방법은 지지도와 신뢰도로 측정될 수 있다. 지지도는 한 규칙이 주어진 데이터집합에 얼마나 자주 적용될 수 있는지를 결정하고, 반면에 신뢰도는 Y에 속한 항목들이 얼마나 빈발하게 X를 포함하여 트랜잭션들에 나타나는지를 결정한다. 주어진 규칙 X→Y에서 신뢰도가 높을수록 X를 포함하는 트랜잭션에 Y가 존재할 가능성이 더 높다. 최소 지지도 임계값과 최소 신뢰도 임계값을 동시에 만족하는 규칙을 강한 규칙이라고 한다. 이러한 측정단위의 정의는 식(7),(8)과 같이 지지도와 신뢰도로 나타낸다.

$$S(X\rightarrow Y)=\frac{\sigma(X\cup Y)}{N} \tag{7}$$

$$C(X\rightarrow Y)=\frac{\sigma(X\cup Y)}{\sigma(X)} \tag{8}$$

4. 실험 및 결과분석

본 논문에서 제안한 연관규칙의 알고리즘을 적용하기 위한 데이터로 UCI 기계학습 데이터 저장소(UCI machine learning repository)의 미국 하원(united states house of representative)의 의원들의 투표 의사록에 연관분석 데이터를 사용하였다. 각 트랜잭션은 한 의원에 대해서 16개 주요의안에 대한 투표기록과 함께 소속정당에 관한 정보를 포함한다. 데이터집합에는 D당(democrat) 267개의 데이터, R당(republican) 168개의 데이터 총 435개의 트랜잭션 데이터와 16개의 항목이 있다. 각 16개의 항목은 이진속성으로 구성이 되어 있으며 1은 찬성, 0은 반대를 의미한다. 또한 데이터 내에 포함되어진 bl는 기권을 의미한다. 전체 트랜잭션데이터베이스에 존재하는 데이터의 속성에는 찬성을 의미하는 1과 반대를 의미하는

0, 그리고 기권을 의미하는 bl이 있다. 모의실험을 구현하기 위한 환경으로는 Windows 7, 64 bit 운영체제에서 MS SQL Server 2012의 데이터처리용 DBMS를 가지고 Visual Studio 개발환경에서 구현하였다. 트랜잭션데이터에서 R당의 투표현황을 Table 1과 같이 나타낼 수 있다.

Table 1 Voting Status of R Party

	x_1	x_2	x_3		x_{14}	x_{15}	x_{16}	Ave
Agreement	31	75	22		158	14	96	83
Objection	134	73	142	...	3	142	50	77
Abstention	3	20	4		7	12	22	8
total	168	168	168		168	168	168	168

식 (3),(4),(5)를 이용하여 R당의 사전확률 값을 계산하면 Table 2와 같이 나타낼 수 있다.

Table 2 Prior Probability values of R Party

	x_1	x_2	x_3		x_{14}	x_{15}	x_{16}	Ave
Agreement	0.19	0.45	0.13		0.94	0.08	0.57	0.49
Objection	0.80	0.43	0.85	...	0.02	0.85	0.30	0.46
Abstention	0.01	0.12	0.02		0.04	0.07	0.13	0.05
Total	1	1	1		1	1	1	1

계산된 사전확률 값과 식(6)을 이용하여 사후확률 값 즉, 가중치를 구한 뒤 최소 가중치($w_{\min}$)인 0.5미만인 규칙을 제거한다. 전체 168개의 규칙 중 최소 가중치를 만족하지 못하는 규칙 25개 규칙이 제거되고, 만족하는 규칙 143개의 규칙만 남게 된다. 같은 방법으로 D당의 전처리를 실행하면 267개 규칙 중 213개의 규칙이 남게 된다. 전처리단계를 거친 감축된 규칙에 T-알고리즘을 이용하여 연관규칙을 감축한다. 데이터는 이진형식으로 되어 있으나 사용된 데이터에는 반대를 의미하는 0과 찬성을 의미하는 1뿐만 아니라 기권을 의미하는 bl이라는 속성이 하나 더 존재한다. 따라서 본 실험에서는 기권표를 1로 간주했을 때와 0으로 간주했을 때, 이 두 가지 경우를 두고 실험하였다. T-알고리즘을 적용한 감축방법에 따른 결과는 Table 3과 같다. Table 3은 D당의 기권표가 1일 경우의 결과를 나타낸 표이다. 실험결과 총 267개의 데이터 중 45개의 데이터로 감축되었다.

Table 3 Final Data of D Party : case $bl=1$

Name	X1	X2	X3	..	X14	X15	X16
R27,R31,R23,R116	1	-	1	··	0	0	-
R207,R157,R75,R55	1	-	1	··	0	0	0
R145,R90,R117,R150	1	-	1	··	0	0	1
R224,R106,R124,R50	1	0	1	··	0	1	-
R170,R56R83,R100	-	0	1	··	0	1	0
R190,R13R45,R08	-	0	1	··	0	1	-
R191,R01,R13,R233	-	0	1	··	0	0	0
R96,R16,R22,R60	1	-	1	··	0	-	1
:	:	:	:	:	:	:	:

Table 4는 기권표가 0일 경우의 결과를 나타낸 표이다. 실험결과 총 267개의 데이터 중 40개의 데이터로 감축되었다.

Table 4 Final Data of D Party : case $bl=0$

Name	X1	X2	X3	..	X14	X15	X16
R103,R06,R14,R10	1	-	1	···	0	0	0
R195,R168,R71,R12	0	-	1	···	0	0	0
R72,R75,R117,R128	1	-	1	···	0	0	1
R160,R216,R75,R181	-	0	1	···	0	1	0
R87,R119,R112,R143	-	0	1	···	0	1	0
R62,R22,R45,R178	-	0	1	···	0	1	-
R196,R155,R12,R102	0	0	1	···	0	0	0
R121,R68,R132,R24	-	1	1	···	0	0	0
R96,R16,R22,R60	1	-	1	···	0	-	1
R160,R216,R75,R102	-	0	1	···	0	1	-
:	:	:	:	:	:	:	:

제안한 알고리즘의 효율을 평가하기 위해 Apriori 알고리즘, T 알고리즘과 비교 평가하였다. Table 5에 나타난 것처럼 WT-알고리즘을 사용했을 때 D당의 경우 기권표를 1로 간주했을 때, 전체 267개의 규칙에서 45개로 감축이 되었고, 기권표를 0으로 간주했을 때, 전체 267개의 규칙에서 40개로 감축되었다. 기권표를 0으로 간주했을 때와 1로 간주했을 때의 감축 평균은 약 85%가 감축되었다. 그리고 R당의 경우 기권표를 1로 간주했을 경우 전체 168개의 데이터 중 34개로 감축되었고, 기권표를 0으로 간주했을 경우 31개로 감축되었다. 감축평균은 약 81%로 감축되었다. 제안한 WT 알고리즘을 사용했을 경우

규칙을 보다 효율적으로 감축할 수 있음을 알 수 있었다.

Table 5 Reduction Ratio Comparison of Final Rules

Apriori algorithm($S_{min=20\%}$)					
D Party			R Party		
bl=1	bl=0	Aver.	bl=1	bl=0	Aver.
103/267	91/267	64%	63/168	52/168	66%
62%	66%		63%	69%	
T-algorithm					
D Party			R Party		
bl=1	bl=0	Aver.	bl=1	bl=0	Aver.
75/267	68/267	74%	47/168	42/168	74%
72%	75%		72%	75%	
WT-algorithm					
D Party			R Party		
bl=1	bl=0	Aver.	bl=1	bl=0	Aver.
45/267	40/267	85%	34/168	31/168	81%
83%	86%		80%	82%	

연관규칙의 성립을 위해 한 예를 들면 WT-알고리즘을 적용했을 때, R당의 기권표를 1로 간주했을 경우 추출된 규칙 전건부 규칙 $x_4=1$, $x_5=1$, $x_{13}=1$, $x_{14}=1$일 때 후건부가 1인 경우 지지도는 88%, 신뢰도는 100%이고, R당의 기권표를 0으로 간주했을 경우 전건부 규칙 $x_4=1$, $x_5=1$, $x_{13}=1$, $x_{14}=1$일 때 후건부가 1인 경우 지지도는 86%, 신뢰도는 100%로 나타났다. 또한 D당의 기권표를 1로 간주했을 경우 추출된 전건부 규칙 $x_3=1, x_8=1, x_9=1$일 때 후건부가 1인 경우 지지도는 84%, 신뢰도는 100%이고, D당의 기권표를 0으로 간주했을 경우 전건부 규칙 $x_3=1, x_8=1, x_9=1$일 때 후건부가 1인 경우 지지도는 80%, 신뢰도는 100%로 나타났다.

T-알고리즘을 적용했을 때, R당의 기권표를 1로 간주했을 경우 추출된 규칙 전건부 규칙 $x_4=1$, $x_5=1$, $x_{13}=1$, $x_{14}=1$일 때 후건부가 1인 경우 지지도는 81%, 신뢰도는 100%이고, R당의 기권표를 0으로 간주했을 경우 전건부 규칙 $x_4=1$, $x_5=1$, $x_{13}=1$, $x_{14}=1$일 때 후건부가 1인 경우 지지도는 76%, 신뢰도는 100%로 나타났다. 또한 D당의 기권표를 1로 간주했을 경우 추출된 전건부 규칙 $x_3=1, x_8=1, x_9=1$일 때 후건부가 1인 경우 지지도는 84%, 신뢰도는 100%이고, D당의 기권표를 0으로 간주했을 경우 전건부 규칙 $x_3=1, x_8=1, x_9=1$일 때 후건부가 1인 경우 지지도는 80%, 신뢰도는 100%로 나타났다. 다음은 Apriori 알고리즘에 대한 평가로 $S_{\min}=20\%$ 이상인 데이터만을 평가하였다. Apriori 알고리즘으로 추출된 규칙 중 기권표를 0으로 간주했을 때와 기권표를 1로 간주했을 경우로 각각 나누어 실험해 보았다. R당의 기권표를 0으로 간주했을 때 즉, $bl=0$일 때의 신뢰도 측정 결과는 $x_4=1$, $x_5=1$, $x_{13}=1, x_{14}=1$일 경우 지지도 82%, 신뢰도 100%로 나타났고, 기권표를 1로 간주 했을 때 $x_4=1$, $x_5=1$, $x_{13}=1, x_{14}=1$ 경우 지지도 73%, 신뢰도 100%로 나타났다. D당의 경우 기권표를 1로 간주 했을 때 $x_3=1$, $x_8=1$, $x_9=1$인 경우 지지도 68%, 신뢰도 100%로 나타났고 기권표를 0으로 간주했을 때 $x_3=1$, $x_8=1$, $x_9=1$일 경우 지지도 62%, 신뢰도 100%로 나타났다.

규칙 감축률을 살펴보면 Apriori 알고리즘은 65%의 감축률로 나타났고 T-알고리즘은 74%의 감축률로 나타났다. 제안한 WT-알고리즘은 83%의 규칙 감축률로 다른 알고리즘보다 높게 나타났다. Apriori 알고리즘의 지지도는 71.25%로 나타났고 T-알고리즘은 80.25%로 나타났으며, 제안한 WT-알고리즘은 84.5%로 나타났다. 제안한 WT-알고리즘을 사용했을 경우 규칙을 보다 효율적으로 감축시킬 수 있었고 지지도도 함께 증가시킬 수 있음을 알 수 있었다.

5. 결 론

광범위한 데이터베이스에서 존재하는 데이터의 종류가 다양해지고 그 양도 폭발적으로 증가하고 있지만 필요한 정보를 찾기가 어려워져 정보의 빈곤현상을 겪고 있다. 따라서 빅데이터 속에 존재하는 유용한 정보를 찾기위한 데이터마이닝의 중요성이 부각되고 있다. 데이터마이닝의 기법 중 가장 많이 사용되고 있는 연관성 분석은 데이

터베이스 내 존재하는 항목들 간의 관계를 기술하는 것으로 규칙의 표현이 간략하며 사용자에게 데이터에 관한 유용한 정보를 줄 수 있다. 그렇지만 이러한 연관성 분석은 항목의 수가 늘어날수록 규칙의 수도 기하급수적으로 늘어나게 되므로 고속처리 방식의 실현이 중요한 과제가 되고 있다.

본 논문에서는 연관규칙 마이닝에서 방대한 데이터베이스에 저장되어 있는 데이터로부터 유용한 정보 및 지식을 추출하는데 가장 잘 알려져 있고, 많은 연구가 이루어지고 있는 연관규칙을 기반으로 하였으며 Boolean 대수 간소화로 알려진 Quine-McKluskey 방법을 알고리즘화 하여 연관규칙 마이닝에 기반한 규칙 감축방법을 제안하였다. 감축방법은 변수의 개수와 관계없이 간략화가 가능한 장점이 있다. 연관규칙을 감축하는 알고리즘의 첫 번째 단계는 전처리단계로서 데이터를 분석에 적합한 형태로 변환하고 중복된 속성들을 제거하는 단계로서 각 항목의 속성에 따라 가중치를 부여하여 항목의 가중치가 낮은 규칙은 제거하는 방법으로 규칙을 감축하였다. 그 다음 단계로 T-알고리즘을 이용하여 변수의 개수에 관계없이 연관규칙을 감축시켜 마이닝 효율을 향상시킬 수 있었다.

제안된 방법의 타당성을 검증하기 위한 모의실험 결과 Apriori 알고리즘은 65%의 감축률로 나타났고 T-알고리즘은 74%의 감축률로 나타났다. 제안한 WT-알고리즘은 83%의 규칙 감축률로 다른 알고리즘보다 높게 나타났다. Apriori 알고리즘의 지지도는 71.25%로 나타났고 T-알고리즘은 80.25%로 나타났으며, 제안한 WT-알고리즘은 84.5%로 나타났다. 제안된 알고리즘은 연관규칙의 감축뿐만 아니라 규칙간의 지지도를 함께 효과적으로 향상 시킬 수 있음을 알 수 있었다. 따라서 연관규칙 탐색방법인 Apriori 알고리즘과 이전에 제안한 T-알고리즘과 비교하였을 때 제안한 알고리즘의 효율성이 높음을 실험을 통하여 확인하였다. 이를 통하여 불필요한 규칙을 감축시켜 데이터마이닝의 효율을 향상시킬 수 있었다.

향후 과제로서 다중치의 개념을 적용한 데이터마이닝의 연관규칙 감축방법이나 다치 속성 값 처리에 대한 연구와 기계학습 알고리즘을 이용하여 소셜네트워크 서비스에서 발생하는 대규모 데이터에 활용할 수 있는 방법이 요구된다.

References

[1] H. C. Park, “Association rule ranking function by decreased lift influence,” Journal of the Korean Data & Information Science Society, Vol. 21, No. 3, pp. 397-405, 2010.

[2] K. C. Ahn, C. B. Moon, B. M. kim, Y. S. Shin, and H. S. kim, “POS Data Analysis System based on Association Rule Analysis,” Korea Society of Industrial Information Systems, Vol. 17, No. 5, pp. 9-17, 2012.

[3] W. Lin, S. Alvarez, and C. Ruiz, “Efficient adaptive-support association rule mining for recommender systems,” Data Mining and Knowledge Discovery, Vol. 6, No. 1, pp. 83-105, 2002.

[4] B. W. Zheng, J. M. Yeo, "The method of using database technology to process rules of Rule-based System," Journal of information and communication convergence engineering, Vol.8, No.1, pp. 89-94, 2010.

[5] J. H. Park, H. M. Chung, "An Effective Reduction of Association Rules using a T-Algorithm," Korea Intelligent Information System Society, Vol. 19, No. 2, pp. 285-290, 2009.

[6] J. H. Park, C. H. Her, H. M. Chung, "Efficiency for Reduction of Association Rule,“ Proceedings of KIIS conference, Vol. 18, No. 2, pp. 101-104, 2008.

[7] Y. Kim, "A Study on Design and Implementation of Personalized Information Recommendation System based on Apriori Algorithm," Journal of the Korean Biblia Society for Library and Information Science, Vol. 23, No. 4, pp. 283-308, 2012.

[8] R. Argrawal, T. Imieliński, and A. Swami, "Mining Association Rules between sets of items in Large Databases," In Proc. Int'l Conf. on Management of Data, ACMSIGMOD, Washington D.C, Vol. 22, No. 2, pp. 207-216, 1993.

[9] R. Srikant, Q. C.Vu and R. Agrawal, "Mining Association Rules with Items Constraints," In Proc. the 3rd Int'l Conf. on Knowledge Discovery and Data mining, pp. 67-73, 1997.

[10] H. S. Hwang, K. D. Yoo, "Mining Association Rules from the Web Access Log of an Online News Website," Korea Society of Industrial Information Systems, Vol. 18, No. 2, pp. 47-57, 2013.

[11] R. Srikant, R. Argrawal, "Mining quantitative association rules in Large Relational Tables," In Proceedings of the ACM SIGMOD Conference on Management of Data, Vol. 25, No. 2, pp. 3-8, 1996.

[12] S. I. Jeon, G. W. Park, K. W. Nam, and K. H. Ryu, "Pattern Analysis-Based Query Expansion for Enhancing Search Convenience," Korea Society of Industrial Information Systems, Vol. 17, No. 2, pp. 65-72, 2012.

[13] K. H. Joo, E. Y. Shin, J. I. Lee ,W. S. Lee, "Hierarchical Automatic Classification of News Articles based on Association Rules," Journal of Korea Multimedia Society, Vol. 14, No. 6, pp. 730-741, 2011.

박 진 희 (Jin-Hee Park)

- 정회원
- 대구가톨릭대학교 컴퓨터정보통신공학과 공학사
- 대구가톨릭대학교 전산통계학과 이학석사
- 대구가톨릭대학교 컴퓨터정보통신공학과 공학박사
- 대구한의대학교 교양학부 조교수
- 관심분야 : 데이터마이닝, 스마트교육

피 수 영 (Su-Young Pi)

- 정회원
- 대구효성여자대학교 전산통계학과 이학사
- 대구가톨릭대학교 전산통계학과 이학석사
- 대구가톨릭대학교 전산통계학과 이학박사
- 대구가톨릭대학교 교양교육원 조교수
- 관심분야 : 데이터마이닝, 인공지능, IT융합