Journal of Digital Contents Society Vol. 17 No. 2 Apr. 2016(pp. 81-88)
http://dx.doi.org/10.9728/dcs.2016.17.2.81

빈발도와 가중치를 이용한 서비스 연관 규칙 마이닝

황정희*

요 약

일반적인 빈발패턴 탐사 방법은 항목의 빈발도만을 고려한다. 그러나 유용한 정보를 추출하는 데 있어 빈발도와 더불어 고려해야 하는 것은 빈발항목이 아니더라도 연관된 항목이 주기적으로 함께 발생한다면 시기나 시간에 따라 관심의 중요도가 변화하는 것을 고려해야 한다. 즉, 시간에 따라 사용자가 요구하는 서비스의 중요도는 다르므로 각 서비스 항목에 대한 중요도의 값을 고려하여 마이닝 하는 방법이 필요하다. 본 논문에서는 서비스 온톨로지 기반으로 가중치를 이용한 서비스 빈발 패턴을 추출하는 마이닝 기법을 제안한다. 제안하는 기법은 시공간 상황을 기반으로 서비스의 중요도를 고려한 가중치를 부여하여 연관 서비스를 발견한다. 새롭게 탐사되는 서비스는 저장되어 있는 서비스 규칙과의 새로운 조합을 통해 사용자에게 최적의 서비스 정보를 제공할 수 있는 기반이 된다.

키워드 : 연관규칙, 데이터 마이닝, 빈발패턴, 온톨로지

Mining Association Rule on Service Data using Frequency and Weight

Jeong Hee Hwang*

Abstract

The general frequent pattern mining considers frequency and support of items. To extract useful information, it is necessary to consider frequency and weight of items that reflects the changing of user interest as time passes. The suitable services considering time or location is requested by user so that the weighted mining method is necessary. We propose a method of weighted frequent pattern mining based on service ontology. The weight considering time and location is given to service items and it is applied to association rule mining method. The extracted rule is combined with stored service rule and it is based on timely service to offer for user.

Keywords : Association rule, Data mining, Frequent pattern, Ontology

1. 서론

유비쿼터스 환경에서 사용자가 원하는 환경을 조성하거나 특정 이벤트를 감지하고 처리하는 등의 서비스 제공에 대한 연구들이 이루어지고 있다. 적합한 서비스를 제공하기 위한 방법에서 서비스 발견과 조합은 중요하다. 유용한 서비스를 제공하기 위해 함께 고려되어야 하는 점은 다양한 사용자들의 서비스 요청에 맞게 사용자의 시간과 위치에 따른 상황정보를 고려하는 동적인 서비스 발견을 통해 사용자에게 적합한 서비스를 제공하는 방법이 필요하다[1, 2].

일반적인 빈발 패턴 마이닝은 각 항목들의 중요도를 고려하지 않고 수행하였으나 실제의 환경에서는 시간에 따른 항목들의 중요도가 다르게 적용되어야 하는 경우가 빈번하다. 그리고 같은 항목이라도 시간에 따라 중요도가 변화하는 경우도 있다. 가중치 패턴 마이닝(Weighted Pattern Mining)은 항목들이 다른 중요도를 고

※ Corresponding Author: Jeong Hee Hwang
Received: November 30, 2015
Revised : February 10, 2016
Accepted : April 26, 2016
* Namseoul University Computer Engineering
Tel: +82-41-581-2108 , Fax: +82-41-581-2100
email: jhhwang@nsu.ac.kr
▣ 이 논문은 2015년도 남서울대학교 학술연구비 지원에 의해 연구되었음

려하여 높은 가중치의 빈발 패턴을 발견하는 마이닝 기법이다. 즉, 같은 서비스 항목이라도 시기에 따른 중요도가 다르므로 다른 가중치를 설정하여 마이닝을 수행한다[5, 6]. 본 논문에서는 서비스에 대한 가중치를 고려하고, 다계층간의 연관 규칙에 기반을 두고 있는 레벨 교차 알고리즘[11, 12]을 이용하여 서비스 빈발 패턴을 탐색하는 방법을 제안한다. 탐색된 빈발 서비스 패턴은 이미 저장된 기존 서비스와의 조합을 통해 사용자에게 적합한 서비스를 제공할 수 있는 기반이 된다. 서로 다른 레벨의 항목에 대한 빈발 항목을 발견할 수 있는 레벨 교차 알고리즘은 더 많은 서비스를 발견할 수 있다. 그러므로 본 논문에서는 시공간 정보 및 서비스의 계층구조를 갖는 온톨로지를 기반으로 하여 서비스 온톨로지의 상위계층에서 하위계층간의 연관탐색을 통해 의미있는 빈발 서비스 패턴의 발견을 가능하게 한다.

기존의 연관 규칙 마이닝 방법은 빈발 이벤트만을 고려하므로 자주 발생하지 않지만 중요도가 높은 이벤트에 대한 연관 규칙은 탐사하지 못한다[4, 5]. [3]에서는 최소 지지도 이하로 발생하지만 특정 데이터와 높은 확률로 함께 발생하는 의미있는 최소 데이터 쌍에 대한 연관 정보 탐사기법을 제안하였다. 제안 방법은 상대 지지도(relative support)에 기반하여 연관 규칙을 탐사하므로 빈발 지지도 중심의 연관 규칙 탐사 방법에서는 탐사할 수 없는 희소 이벤트 쌍에 대한 연관 규칙을 탐사한다. [1]에서는 패턴의 중요도에 따라 가중치를 부여하여 패턴 사이에 존재하는 연관 규칙을 탐사하는 Weighted Interesting Patterns 방법을 제안하였다. 이 방법은 패턴의 중요도 및 관심도에 따라 가중치를 높게 부여함으로써 보다 중요한 패턴에 대한 연관 규칙을 탐사할 수 있다.

가중치를 두는 패턴 마이닝(Weighted Pattern Mining)은 항목들이 다른 중요도를 가질 경우를 고려하여 높은 가중치 패턴을 찾아내는 마이닝 기법을 의미한다. 비즈니스 데이터 분석 환경에서 상품에 대한 고객들의 구매 패턴은 시기와 계절에 따라 다르게 설정되어야 하고, 다양한 상품 가격에 대해서도 다른 가중치가 설정되어야 한다. 가중치 빈발 패턴 마이닝 알고리즘은 초기에 Apriori 알고리즘을 기반으로 하는 WIP[2], WARM[6], WAR[7] 등이 있다. 그러나 이러한 알고리즘들은 여러 번의 데이터베이스 스캔을 필요로 하여 속도가 느려지는 단점을 가지고 있다. 이러한 성능상의 문제를 해결하기 위한 방법으로 WFIM[8]이 제안되었다. [8]에서는 항목들의 최소 가중치와 가중치 범위를 정하고 FP-트리를 가중치 오름차순으로 구성하여 FP-트리가 하향 닫힘(Downward Closure)성질을 만족하도록 한다. WIP[2] 알고리즘은 가중치 패턴의 성질을 Weight Affinity 개념을 이용하여 정의하고 Weight Affinity 값이 큰 패턴을 유용한 패턴으로 하여 탐색하고 있다. 이들 WFIM와 WIP는 FP-Growth 알고리즘과 같이 두 번의 데이터베이스 스캔을 필요로 한다.

일반적인 순차패턴 마이닝에서는 분석 대상 데이터 집합에 포함되는 구성요소의 발생 순서만을 고려한다. 그러므로 단순한 순차패턴은 쉽게 찾을 수 있는 반면 실제 응용 분야에서 널리 활용될 수 있는 관심도가 큰 순차패턴을 탐색할 수 없는 단점이 있다. 이를 보완하는 가중치 순차패턴 탐색[9, 10]에서는 관심도가 큰 순차패턴을 얻기 위해서 구성요소의 단순 발생 순서뿐만 아니라 실제 응용분야에서 단위 항목에 대한 서로 다른 중요도를 고려하여 흥미도나 관심도가 큰 순차패턴을 탐색할 수 있는 방법을 제시한다.

본 논문에서는 사용자에게 최적의 서비스를 제공하기 위한 방법으로 가중치 기반의 서비스 온톨로지를 이용하여 서비스 빈발 패턴을 추출하는 마이닝 기법을 제안한다. 온톨로지를 기반으로 마이닝을 수행하는 것은 특정 도메인 개념을 다계층으로 고려하여 데이터 마이닝을 적용하는 것과 유사한 방법이다. 같은 레벨간의 연관된 정보만을 추출하면 같은 레벨에서 연관된 정보가 없으면 추출되지 않는 문제가 발생할 수 있다. 그러므로 다른 레벨의 항목이라도 연관된 항목이 존재하면 발견할 수 있는 계층간의 빈발도를 함께 고려하는 알고리즘을 적용한다. 같은 서비스 항목이라도 시간에 따라 가중치가 다르게 적용되어 사용자에게 제공되는 서비스 항목이 달라질 수 있으므로 서비스의 질에도 영향을 미친다.

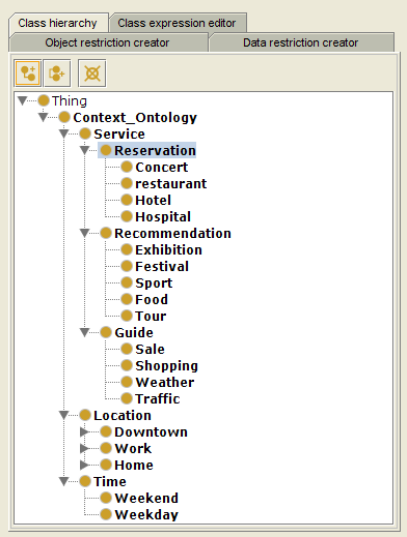
2. 온톨로지와 마이닝

사용자의 환경과 배경을 개념적인 도메인으로 온톨로지를 구성하는 컨텍스트 온톨로지는 시간과 공간 그리고 사용자의 행동에 대한 정보를 모두 포함하므로 정확한 서비스를 제공할 수 있는 기초가 된다. 즉, 미리 정의된 컨텍스트 온톨로지 스키마를 사용하여 개인화된 서비스를 상황에 따라 언제 어디서나 제공할 수 있다. 본 논문에서는 상황정보를 구조화하는 데 유용하고 상호 관계성 및 부분적인 상황의 정보를 표현할 수 있는 온톨로지를 설계한다. 설계된 온톨로지는 시간과 공간, 그리고 서비스 정보로 구성되며 이는 사용자의 상황정보 서비스와 연관된다.

사용자의 위치 정보는 실제 이벤트가 발생한 유효시간에 대한 공간상의 좌표에 대한 일반화 값으로 구성된다[1]. 시간은 일정한 시간간격으로 구분한 정보이고, 공간정보는 공간상의 좌표 (x_i, y_i)에 해당하는 일반화된 위치 정보이다. 위치 일반화는 공간의 좌표 값에 대해 일정한 구역(zone)으로 일반화하여 사용자 및 객체의 위치를 식별한다. 그러므로 시공간 정보는 사용자의 상황정보의 시점을 나타내고 시간에 따른 서비스 이력을 구분할 수 있게 하는 기준이 된다.

(그림 1)은 상황정보를 표현하는 온톨로지 구조를 protege을 이용하여 구성한 것이다. 사용자를 위한 서비스는 시간과 공간 정보가 연관된 서비스 규칙으로 구성된다. 시간 정보는 일상적인 생활의 행동패턴으로 나누어 구분할 수 있는 서비스 제공의 기준이 되는 주중, 주말 정보로 구분하고 이에 대한 하위레벨로 오전, 오후, 야간으로 나뉘어 일반화하였다. 위치 정보는 사용자들의 일상생활에서 가장 많은 시간을 보내게 되는 장소인 홈, 직장, 다운타운으로 구분하여 일반화하고 해당 장소에서 많은 사람들이 이용하는 공간으로 다시 세분화하였다. 홈 공간은 방, 부엌, 거실로, 직장은 사무공간, 로비, 카페테리아 등으로 세분화한다. 서비스는 가이드, 추천, 예약으로 구분되며, 가이드는 다시 하위 레벨의 교통, 날씨 등으로, 추천서비스는 상품 추천 및 숍 추천 등으로, 예약서비스는 호텔예약 및 운송수단예약 등으로 세분화하였다.

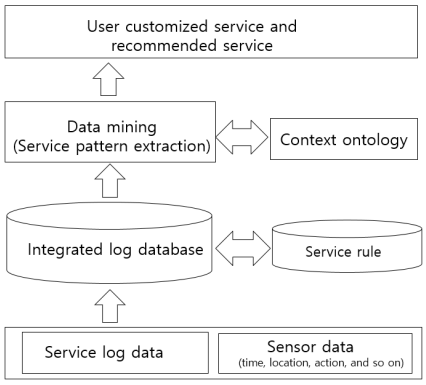
(그림 1)서비스 상황정보 온톨로지



(Figure 1) Service context ontology

사용자 요청 서비스 및 추천 서비스를 추출하는 과정을 (그림 2)에서 보여준다. 시공간 정보를 포함한 사용자의 행위에 따른 기본적인 센서 정보와 서비스 사용 이력 정보는 통합 데이터베이스에 저장된다. 통합 데이터는 온톨로지에 저장되어 있는 서비스 규칙과의 조합을 통해 새로운 서비스 규칙을 발견하는 기초가 되고, 연관된 빈발 서비스 패턴을 추출하는 데이터 마이닝을 수행한다. 마이닝을 통해 추출된 서비스 연관규칙은 사용자에게 필요한 서비스를 시공간 정보를 고려하여 사용자를 위한 맞춤 서비스와 추천을 위한 서비스 정보를 제공한다.

(그림 2) 서비스 제공 과정



(Figure 2) Service process

3. 가중치를 이용한 마이닝 알고리즘

가중치를 고려하면 빈발하지 않더라도 상대적으로 중요한 이벤트에 대한 연관규칙을 탐사할 수 있다. 즉, 가중치를 이용하기 때문에 더 많은 규칙 및 시기를 반영한 규칙 발견이 가능하다. <표 1>은 온톨로지에 저장된 서비스 규칙의 예를 보여준다. 사용자의 기본 정보와 함께 시간과 장소에 따라 사용자가 자주 사용하는 서비스 규칙을 저장한다. 여기서 서비스 규칙 Service1, service2는 서비스 온톨로지 구조에 따라 부여된 특정 서비스 식별자를 포함한다.

<표 1> 온톨로지 서비스 규칙

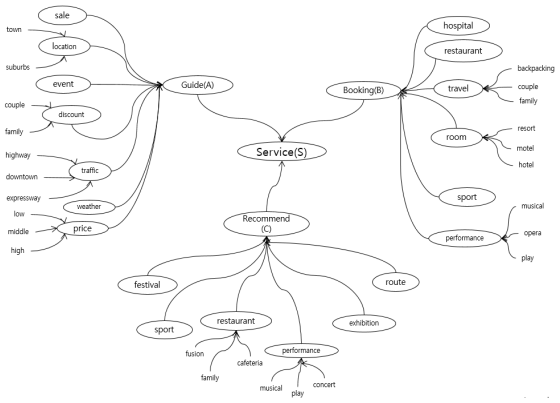
UserID	Time	Location	Service1	Service2
U01	Tc1	Lb3	Sa4	Sd2
U02	Ta2	La2	Sc3	Sd1
U03	Tb2	Lt1	Sb2	Sc3
U04	Td1	Lc2	Sd1	Sf2

<Table 1> Ontology service rule

서비스 빈발 패턴의 의미는 해당시간과 위치에서 해당 서비스가 자주 이용된다는 것을 말한다. 예를 들어, 빈발 서비스 패턴 (Sa3, Sc1, Sd1)은 추가적으로 (Sa3, Sc1), (Sa3, Sd1), (Sc1, Sd1)도 빈발 서비스 패턴이 된다. 이러한 저장된 서비스 규칙외에 추가적으로 기준이 되는 일정한 시간정보 또는 위치정보에 대한 서비스 계층내에서도 관심있는 조건의 추가적인 연관규칙을 발견할 수 있다.

본 논문에서는 마이닝 과정을 예로써 설명하기 위하여, 일정한 시간과 공간 정보에 대한 서비스 사용 트랜잭션에 대한 정보를 가지고 마이닝을 수행한다고 가정한다. 그리고 최소 지지도와 최소 가중치를 이용하여, 서비스 온톨로지에 대한 계층정보를 사용하여 마이닝을 수행하는 과정을 설명한다. 서비스 레벨은 3레벨로 정의되며 하위 레벨은 부모 레벨의 서비스 항목에 대한 상세한 서비스 항목을 의미한다. 서비스 레벨에서 하위 레벨은 부모 레벨의 가중치보다 작거나 같은 가중치를 갖는다.

(그림 3) 서비스 온톨로지



(Figure 3) Service ontology

(그림 3)의 서비스 정보에 대한 단순한 표현을 위해 첫 번째 레벨은 대문자알파벳으로, 두 번째 레벨은 소문자알파벳으로, 세 번째 레벨은 숫자로 트랜잭션의 항목을 나타낸다. 서비스 정보에 대한 알파벳 표기방법은 같은 레벨의 서비스에 대해 순차적으로 알파벳을 부여한다 즉, 1-레벨은 대문자 알파벳 표기로써 안내(A)-예약(B)-추천(C) 식으로 표기하고, 같은 부모를 같은 2레벨의 서비스 내역에는 순차적인 소문자 알파벳을 부여한다. 안내 서비스의 위치(a)-할인(b)-이벤트(c)-교통(d)-세일(e)-가격(f) 등으로 부여하고, 예약 서비스의 숙박(a)-스포츠관람(b)-여행(c)-공연(d)-레스토랑(e)-병원(f) 등으로 부여한다. 또한 이들의 하위계층은 숫자 1부터 차례로 부여한다. 예를 들어, 항목 Aa2는 안내서비스로 관광지 위치에 따른 안내서비스를 의미한다.

<표 2> 트랜잭션 데이터

TID	Items
100	Aa2, ~~Bc1~~, Cc1, Ba1, ~~Cd1~~, ~~Af1~~
200	Ad3, Bd2, Cc1, Ba1, Ab1
300	~~Cf2~~, Aa1, ~~Af2~~, ~~Ae1~~. Ba1, Bd1
400	Cc2, Bd2, Ad2, ~~Bb1~~, Ab1, ~~Be1~~
500	Cc1, Ab1, Ad1, Aa2, Bd2

<Table 2> Transaction Data

<표 2>는 서비스 이용에 대한 트랜잭션 데이

터의 예이다. 빈발 데이터 항목의 추출을 위해 최소 빈발도(min_sup) 3와 최소 가중치(min_wght) 1.5로 가정하여 적용한다. 데이터에 따라 빈발도를 제외하고 가중치만을 고려하여 적용할 수도 있다. 빈발도를 고려한 항목의 가중치는 빈발도와 가중치가 함께 계산한다. 즉, 항목의 가중치는 항목별 세부가중치*빈발도로 계산하여 적용한다.

<표 3> 필터링된 2-level 항목

Itemset	Wsupport
Aa*	2.4(0.8 *3)
~~Ab*~~	~~1.2(0.4 *3)~~
Ad*	2.4(0.8 *3)
Ba*	2.7(0.9 *3)
Bd*	2.0(0.5 *4)
Cc*	3.2(0.8 *4)

<Table 3> Filtered 2-level Items

<표 4> 필터링 데이터

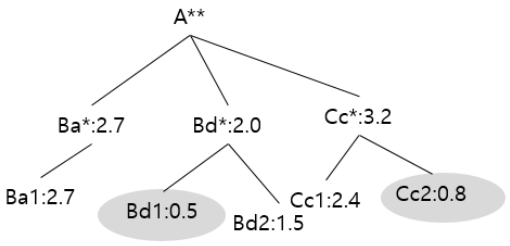
TID	Items
100	Aa2, Ba1, Cc1
200	Ad3, Bd2, Cc1, Ba1
300	Aa1, Bd1, Ba1
400	Ad2, Bd2, Cc2
500	Aa2, Ad1, Bd2, Cc1

<Table 4> Filtered data

<표 2>의 트랜잭션 데이터에서 1레벨의 A, B, C 기준으로 1레벨 중심의 1차적인 빈발 항목을 필터링한다. 즉, A**, B**, C**에 대한 빈발도를 기준으로 후보항목을 필터링한다. 그러나 <표 2>의 데이터는 1레벨의 종류가 A, B, C로 적으며, 최소 빈발도 3을 모두 충족하므로 다음 레벨인 2레벨 기준으로 빈발도를 만족하지 않는 것을 밑줄로 표시한 것이고, 최소 빈발도를 만족하는 항목만을 필터링하여 <표 3>에서 보여준다. <표 3>에서 항목별 가중치를 계산하면, 빈발도 3을 만족하는 항목중에서 Ab*:3는 빈발도는 만족하지만 항목의 세부 가중치가 0.4로 항목의 가중치는 1.2(가중치*빈도: 0.4*3)이므로 2차적인 필터링 과정에서 제거된다. 그러므로 <표 2>의 원시 데이터에서 필터링한 결과는 <표 4>이다.

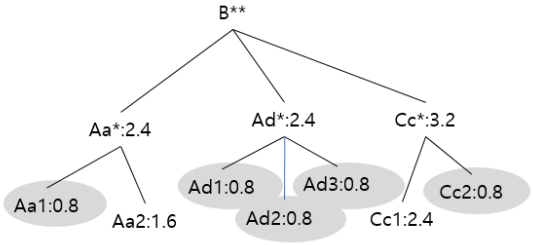
다음은 <표 4>의 데이터에서 각 항목기준으로 빈발항목간의 연관 빈발 패턴을 추출한다. 주어진 최소 지지도와 최소 가중치를 만족하는 항목의 하위 레벨도 함께 고려하여 A**항목에 대한 레벨 교차트리를 (그림 4)에서 보여준다. Bd1과 Cc2는 최소 임계치를 만족하지 않아 빈발항목에서 제외되므로 결과적으로 A**항목 기준의 빈발항목은 A**-Ba1-Bd2-Cc1이 된다.

(그림 4) A**기준의 교차 빈발 트리



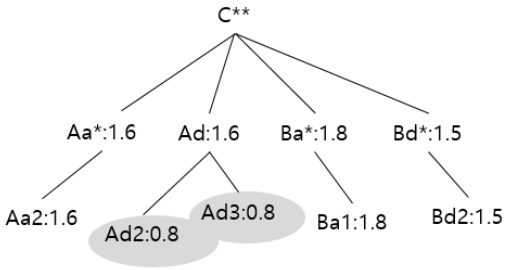
(Figure 4) A** based cross-level tree

(그림 5) B** 기준의 교차 빈발 트리



(Figure 5) B** based cross-level tree

(그림 6) C** 기준의 교차 빈발 트리



(Figure 6) C** based cross-level tree

이와 같은 방법으로 (그림 5)와 (그림 6)에서 B**와 C**에 대한 레벨 교차항목관계를 나타내며, B**항목에 대한 빈발항목은 B**-Aa2-Ad*-Cc1이고, C**항목에 대한 빈발항목은 C**-Aa2-Ad*-Ba1-Bd2으로 나타난다.

<표 5>는 A**, B**, C**기준으로 최소 가중치를 만족하여 발견된 빈발항목을 나타낸 것이다.

<표 5> 레벨교차 빈발 연관 서비스 항목

Frequent Association Items		
Level1	Level2	Level3
A**	Ba*:2.7, Bd*:2.0, Cc*:3.2	Ba1:2.7, Bd2:1.5, Cc1:2.4
B**	Aa*:2.4, Ad*:2.4, Cc*:3.2	Aa2:1.6, Cc1:2.4
C**	Aa*:1.6, Ad*:1.6, Ba*:1.8, Bd*:1.5	Aa2:1.6, Ba1:1.8, Bd2:1.5

<Table 5>Cross-level based frequent service association rules

위에서 연관규칙을 추출하는 과정을 알고리즘으로 기술하면 다음과 같다.

```
Input: service transaction table, T[1],
       min_sup, min_wght
Output: cross-level frequent service tree
If i_level =1  {
   If F[i_level, 1_item] ≥ min_sup
     and W[i_level, 1_item] ≥ min_wght
     insert frequent_service_item
            into filtered table, T[2]
   }
//Extract frequent items by levelling up
For (i_level=2; i_level≤ max_level; i_level++) {
  For (j=1; j≤transaction_cnt; i++)   {
     If  F[i_level, j_item ] ≥ min_sup
        and W[i_level, j_item] ≥ min_wght
        make  cross_level  frequent_item_set;
 }
}
```

알고리즘에 의해 생성된 교차 레벨 빈발트리에서 최대의 빈발 항목관계를 발견할 수 있다. A**항목에 대한 빈발항목은 B**와 C**항목 기주의 빈발트리에서 공통으로 발생하는 A**항목을 찾고, B**항목에 대한 빈발항목은 A**와 C**에서의 B** 공통 빈발항목을, C**항목에 대한 빈발항목은 A**와 B**에서의 공통 빈발항목을 발견하면 된다. 즉, A**에 대한 빈발항목은 B**와 C**항목과 함께 발생하는 B**-Aa2-Ad*-Cc1, C**-Aa2-Ad*-Ba1-Bd2에서의 공통항목인 Aa2, Ad*가 된다. 그리고 B**의 빈발항목은 A**와 C**항목에서 함께 발생하는 Ba1, Bd2가 된다. 이와 같은 방법으로 C**의 빈발 항목은 Cc1이 된다. 따라서 최대 빈발항목 서비스는 Aa2-Ad*-Ba1-Bd2-Cc1(관광지 위치안내-교통안내-숙박콘도예약-연극공연예약-퓨전맛집추천)이 된다. 공통 빈발항목을 발견하는 과정에서 Aa2, Aa가 있다면 공통 항목은 Aa가 된다. 동일 카테고리 서비스에서 상위레벨은 하위레벨의 상세 정보를 포함하므로 사용자의 관심도에 따라 레벨을 조절하여 빈발 항목을 발견할 수 있다. 그리고 연령별, 시기별, 계절별에 따라 가중치를 다르게 하여 정보를 추출하면 더 유익한 서비스 정보를 제공할 수 있다.

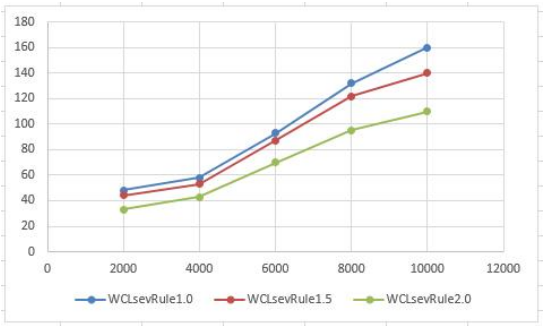
4. 실험

이 장에서는 제안된 가중치 기반 마이닝 알고리즘의 성능 평가를 위해 가중치의 임계치를 다르게 하여 수행 속도 및 발견되는 규칙의 수를 비교하였다. 그리고 가중치를 고려하여 규칙을 발견하는 경우의 효율성에 대해 기술한다. 실험에서는 10,000개의 트랜잭션을 대상으로 한다. 각 트랜잭션에는 랜덤하게 2~3개의 기본 서비스 정보를 포함한다. 실험에서 트랜잭션을 2,000개 단위로 순차적으로 입력하여 수행시간 및 규칙 수의 변화를 측정하였다.

첫 번째 실험에서는 가중치의 변화(1.0, 1.5, 2.0)에 따른 WCservRule_1.0 WCservRule_1.5, WCservRule_2.0의 수행시간을 비교하였다. 그림에서 보는 것처럼 가중치가 커질수록 수행시간이 적게 걸리는 것을 볼 수 있었다. 이것은 가중치가 높으면 임계치를 충족하지 않아 필터링되

는 항목 즉, 빈발 후보항목이 줄어들기 때문에 전체적인 수행시간이 줄어드는 것이다.

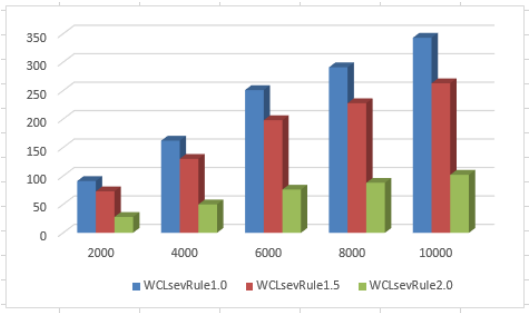
(그림 7) 트랜잭션 증가에 따른 수행시간



(Figure 7) Execution time

두 번째 실험에서는 가중치의 변화에 따른 빈발 항목 수의 변화를 통해 생성되는 규칙 수의 변화를 알아보는 실험을 하였다. 가중치가 커질수록 발견된 빈발 항목 수가 적어지는 것을 볼 수 있었다. 빈발 항목 수의 계산은 빈발항목 Aa1의 경우 A, Aa, Aa1도 빈발 항목 수에 포함되는 것으로 계산하였다. 빈발 항목의 수가 적어지면 조합 가능한 규칙의 수가 적어진다는 것을 의미를 포함한다. 한편 빈발 항목이 많이 발견되는 것만이 우선적으로 좋다고 판단할 수도 없다. 발견된 규칙의 정확성과 유용성을 고려해야 하기 때문이다. 그러므로 양질의 정보를 발견하기 위해서는 데이터의 분포와 특성을 고려한 반복적인 실험을 통해 적절한 가중치를 부여하는 것이 중요하다.

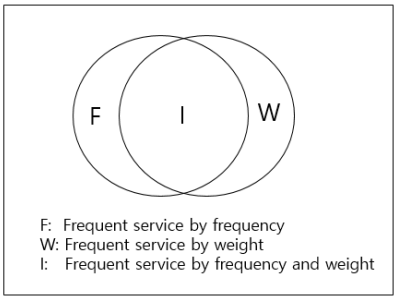
(그림 8) 트랜잭션 증가에 따른 빈발규칙의 수



(Figure 8) The number of frequent rules

온톨로지 규칙 베이스에는 특정 시간 및 위치에 따라 제공가능한 서비스 조합의 규칙이 저장되어 있다. 그러므로 온톨로지에 이미 저장된 서비스 규칙이외에 새롭게 발견되는 서비스 규칙을 추가적으로 조합하여 더 양질의 서비스를 사용자에게 제공하는 것이 가능하다는 것을 알 수 있다.

(그림 9) 빈발 서비스 추출 영역



(Figure 9) Frequent service extraction scope

가중치를 고려하여 빈발 연관규칙 서비스를 발견할 때의 효율성을 설명하기 위한 (그림 9)를 살펴보면 F영역은 빈도만 고려할 때 발견되는 빈발항목을, W영역은 가중치만 고려할 때의 빈발항목 영역을 나타낸다. 그리고 I영역은 빈발도와 가중치를 모두 고려할 때에 발견되는 빈발항목 영역을 나타낸 것이다. 빈발도 또는 가중치만 고려하면 발견되지 않는 항목도 있다는 것을 알 수 있고, 특히 I영역의 항목중에서는 빈발도는 작지만 가중치에 의해 발견되는 희소항목이 포함되는 경우도 있고, 빈도와 가중치를 함께 고려할 때 F와 W영역에서 발견되지 않았던 항목들을 발견할 수도 있다. 그러므로 사용자의 관심도, 데이터의 특성과 분포, 특정 요소에 따른 중요도를 고려하여 빈도와 가중치를 적절하게 적용하는 것이 중요하다. 이와 더불어 분석에 사용될 데이터의 분류도 중요한 요소이다. 각 서비스 항목에 대한 분류를 할 때 너무 세부적으로 분류를 하면 항목의 종류가 많아지게 되어 유용한 연관규칙을 찾기 어려울 수도 있다. 그러므로 서비스 항목별 분류는 되도록 포괄적이고, 일반적인 기준으로 계층화하고 분류하는 것이 양질의 규칙을 발견하기 위한 기초가 될 수 있다. 또한 희소성 있는 항목이나 적은 빈도의 중요도가 있는 항목은 보다 일반화하여 레벨을 상향조정하

면 좋은 결과를 생성할 수 있다.

5. 결론

본 논문에서는 빈발항목이 아니더라도 연관된 항목이 주기적으로 함께 발생하는 것을 발견하기 위하여 항목의 중요도를 고려하는 마이닝 방법을 제안하였다. 시간에 따라 사용자가 요구하는 서비스의 중요도가 달라질 수 있으므로 빈발도만 고려하는 마이닝과는 다르게 항목에 부여된 중요도를 함께 고려하면 빈발도만을 고려하는 마이닝 결과와는 다른 새로운 연관규칙의 항목들이 발견되는 것을 실험을 통해 알 수 있었다. 제안하는 기법은 사용자의 요구 변화와 더불어 시공간 상황을 함께 고려하므로 최신의 정보를 사용자에게 서비스하기 위한 응용에 활용될 수 있다.

References

[1] D. Han, D. Kim, J. Kim, C. Na, B. Hwang, “A Method for Mining Interval Event Association Rules from a Set of Events Having Time Property,” Journal of Korea Information Processing Society, Vol.16-D, No.2, pp.186-190, 2009

[2] U. Yun, J. J. Leggett, "WIP:mining Weighted Interesting Patterns with a strong weight and/or support affinity," SIAM International Conference on Data Mining, pp. 624-628, 2006

[3] H. Yun, D. Ha, B. Hwang, K. Ryu, “Mining Association Rules on Significant Rare Data Using Relative Support,” Journal of Systems and Software, Vol.67, No.3, pp.181-191, 2003

[4] R. J. Swargam, and M. J. Palakal, “The Role of Least Frequent Item Sets in Association Discovery,” In Proc. of International Conference on Digital Information Management, 2007

[5] C. F. Ahmed, S. K. Tanbeer, B. S. Jeong, Y. K Lee, “Mining Weighted Frequent Patterns in Incremental Databases,” Proc. of the Pacific Rim, 2008

[6] F. Tao, “Weighted Association Rule Mining using Weighted Support and Significant Framework,” Proc. of the ACM SIGKDD, 2003

[7] W. Wang, J. Yang, P. S. Yu, “WAR:Weighted Association Rules for Item Intensities,” Knowledge Information and Systems, 2004

[8] U. Yun, J. J. Leggett, “WFIM:Weighted Frequent Itemset Mining with a Weight Range and a Minimum Weight,” Proc. of the Fourth SIAM Int. Conf. on Data Mining, 2005

[9] S. Lo, “Binary Prediction based on Weighted Sequential Mining Method,” Proc. of the Int’l Conf. on Web Intelligence, pp.755-761, 2005

[10] U. Yun, “A New Framework for Detecting Weighted Sequential Patterns in Large Sequential Databases,” Knowledge-Based Systems, 2008

[11] R. S. Thakur, R.C. Jain and K. R. Pardasani, "Mining Level-Crossing Association Rules from Large Databases," Journal of Computer Science 2(1), pp. 76-81, 2006.

[12] V. Ramana, M. Rathnamma, A. Reddy, "Methods for Mining Cross Level Association Rule In Taxonomy Data Structures," International Journal of Computer Applications, Vol. 7, No. 3, 2010.

황 정 희

2001년 :충북대학교 전자계산학과 (이학석사)
2005년 :충북대학교 전자계산학과 (이학박사)

2001년~2006년: 정우시스템(주) 연구소장
2006년~현 재 : 남서울대학교 컴퓨터학과 조교수
관심분야 : 유비쿼터스 컴퓨팅, 데이터 마이닝, 빅데이터