Journal of Digital Contents Society Vol. 17 No. 5 Oct. 2016(pp. 359-366)
http://dx.doi.org/10.9728/dcs.2016.17.5.359

동적 가중치 기반의 연관 서비스 탐사 기법

황정희*

요 약

유비쿼터스 환경에서 사용자에게 유용한 서비스를 제공하기 위해서는 시간과 공간을 기반으로 사용자의 행동과 선호 패턴을 고려하여 가장 적합한 데이터를 처리할 수 있는 방법이 필요하다. 실세계에서 사용자의 관심은 시간이 지남에 따라 변화할 수 있다. 그러므로 서비스 관심도의 변화를 중요도에 반영하여 정보를 추출할 수 있는 방법이 필요하다. 이 논문에서는 사용자에게 필요한 서비스 정보를 온톨로지로 설계하고 시간에 따라 동적으로 변화하는 사용자의 서비스 이용 패턴이나 데이터의 중요도를 동적 가중치로 표현하여 빈발 패턴을 찾는 방법을 제안한다. 이 논문에서 제안하는 동적 가중치를 고려하는 빈발 서비스 패턴 마이닝 기법은 시간의 변화에 따라 필요로 하는 사용자의 관심을 서비스의 중요도로 반영하므로 실시간의 최적화된 서비스 제공이 가능하다.

키워드 : 데이터 마이닝, 가중치 마이닝, 연관 규칙, 패턴 마이닝

An associative service mining based on dynamic weight

Jeong Hee Hwang*

Abstract

In order to provide useful services for user in ubiquitous environment, a technique that can get the helpful information considering user activity and preference is needed and also user's interest actually changes as time passes. Therefore, the discovering method which reflects the concern degree of service information is needed. In this paper, we present the finding method of frequent pattern with dynamic weight on individual item based on service ontology we design. Our method can be applied to provide interested service information for user depending on context.

Keywords : Data mining, Weight Mining, Association rule, Pattern Mining

1. 서론

데이터 마이닝은 대용량의 데이터베이스에서 의미 있는 패턴과 규칙을 발견하고 분석한다. 여러 가지 기법 중에서 데이터로부터 숨겨진 패턴을 추출하는 연관규칙(association rule)은 장바구니 분석이라고 불리기도 하며, 대용량 데이터베이스에서 단위 트랜잭션당 동시에 발행할 확률이 높은 항목들의 유형을 발견하는 기법이다. 이러한 연관규칙은 고객 데이터베이스로부터 구매 품목들간의 관련성을 발견하여 교차 판매 또는 상품 진열에 이용되거나, 우량고객에 대한 상품 카탈로그 발송의 direct-mailing에 이용될 수 있다[1]. 빈발패턴 마이닝은 연관규칙을 발견하거나 데이터간의 관련성을 파악하는 기법으로, 트랜잭션 데이터베이스에 나타난 여러 패턴들 중에서 빈도수가 주어진 임계값보다 크거나 같은 패턴을 찾는 것이다.

온톨로지는 공유된 개념화(shared conceptualization)에 대한 정형화되고 명시적인

※ Corresponding Author: Jeong Hee Hwang
Received : August 20, 2016
Revised : October 15, 2016
Accepted : October 20, 2016
* Namseoul University Computer Engineering
Tel: +82-41-581-2108 , Fax: +82-41-581-2100
email: jhhwang@nsu.ac.kr
▣ 이 논문은 2016년도 남서울대학교 학술연구비 지원에 의해 연구되었음

명세이다. 즉, 온톨로지는 해당 영역의 개념과 이들 개념들 사이의 관계를 설정하여 컴퓨터가 상황을 이해하고 데이터를 해석하는 데 도움을 줄 수 있다. 따라서 관심있는 도메인의 일반적인 지식의 체계를 온톨로지를 통해 상호연관성을 기술하고 이를 이용하면 효율적인 정보 추출이 가능하다. 유비쿼터스 환경에서 사용자의 서비스에 대한 선호도는 상황에 따라 변화하기 쉽다. 사용자에게 유용한 서비스를 제공하기 위해서는 시간과 공간을 기반으로 사용자의 행동과 선호 패턴을 고려하여 가장 적합한 데이터를 처리할 수 있는 시스템이 필요하다. 시스템을 구성하는 가장 중요한 요소는 사용자가 유용하게 참조할 수 정보를 추출하는 방법의 연구가 중요하다. 온톨로지를 이용한 마이닝 방법은 정보를 세분화하여 사용자의 서비스 정보 이용 패턴의 흐름을 추출할 수 있는 기법이다. 그러나 기존의 온톨로지 기반의 마이닝 방법들은 정보의 관심도를 배제한 정보 자체의 형태만을 온톨로지화하여 참조한 마이닝 알고리즘을 통해 유용한 정보를 추출하는 것에 중점을 두고 있다. 실세계에 존재하는 데이터의 중요도는 시간이 지남에 따라 변화할 수 있다. 그러므로 정보의 영향력을 의미하는 사용자의 관심을 중요도에 반영하여 유용한 정보를 추출할 수 있는 방법이 필요하다.

이 논문에서는 빈발패턴 마이닝 중에서 가중치를 고려하는 마이닝 방법을 이용한다. 가중치 패턴 마이닝(Weight Pattern Mining)은 항목들이 다른 중요도를 가질 경우 높은 가중치를 찾아내는 마이닝 기법이다[4-10]. 예를 들면 상품에 대한 고객들의 구매 패턴이나 선호도가 시간에 따라 달라지듯이 사용자가 요구하는 서비스의 관심에 따라 다른 가중치로 설정되어야 한다. 모든 서비스에 대해 동일한 가중치를 적용하여 서비스 빈발 패턴을 발견하여 제공하는 것은 사용자가 요구하는 서비스에 적절하게 적용하지 못하는 경우일 수도 있다. 즉, 계절적 특징이나 휴가철, 연휴 등과 같은 시간의 흐름에 따라 요구되는 서비스의 종류별로 중요도가 달라질 수 있으므로 이것을 서비스 제공 규칙에 반영할 필요가 있다. 사용자의 서비스 이용 패턴은 연령과 계절 및 특별한 기간에 따라 달라질 수 있다. 따라서 이 논문에서 제안하는 동적 가중치를 고려하는 빈발 서비스 패턴 마이닝 기법은 시간의 변화에 따라 필요로 하는 서비스의 중요도가 달라져야 하는 실제적인 사용자의 상황을 반영하는 방법으로써 서비스 정보 제공에 기반이 된다.

2. 관련연구

유비쿼터스 환경에서 발생하는 정보의 홍수 속에서 사용자에게 필요한 정보를 구별하기란 무척 어려운 작업이 되었다. 많은 정보로부터 적절한 정보의 분배와 필터링은 정보 수집 못지 않은 노력이 필요하게 되었다. 이러한 환경에서 상황인지 시스템은 사용자의 정보선택을 도와주고 나아가 맞춤형서비스를 제공하므로 생산된 정보의 효용성을 극대화시킨다. 즉, 사용자가 원하는 정보를 찾아보도록 하는 것이 아니라 사용자의 상황에 맞게 알맞은 정보를 선별하여 제공하게 되므로 사용자의 정보 선택에 많은 영향을 준다.

대량의 데이터로부터 숨겨져 있는 연관된 패턴을 찾아 유용한 정보를 추출하기 위한 기술인 데이터 마이닝은 의사결정에 직접적인 영향을 미치기 때문에 연관관계, 분류, 순차패턴, 특징추출 등을 포함하여 많은 기법이 사용되고 있다. 이러한 기법중에서 연관규칙의 빈발 패턴 탐색 방법이 많이 연구되고 있다. 빈발패턴 마이닝은 트랜잭션 항목에서 미리 정의된 최소 지지도를 만족하는 빈발 항목집합을 찾아내고, 이들 빈발 항목집합들 간의 연관성 정도를 반영하는 연관규칙을 찾아내는 것이다. 즉, 트랜잭션 데이터베이스에서 나타난 여러 패턴중에서 빈도수가 주어진 임계값을 만족하는 패턴을 발견하는 방법이다. Apriori 알고리즘을 기반으로 하는 빈발패턴 마이닝에서 만일 어떤 패턴 a가 빈발하지 않은 패턴이면 a의 모든 슈퍼셋(Super Set)은 빈발하지 않은 패턴이 된다. 이를 Anti-monotone 성질이라고 한다[2].

일반적인 연관규칙 알고리즘에서는 모든 항목에 대한 중요성을 동일한 것으로 간주한다. 그러나 실제 응용 분야에서 단위 항목들은 서로 다른 중요성을 가진다. 그러므로 빈발패턴 탐사에서 단위 항목에 대한 차별화된 중요성을 고려하는 경우 사용자의 흥미도나 관심도가 큰 서비스 정보를 얻을 수 있다. 가중치 빈발 패턴 마이닝에서 단위 항목의 가중치는 연령이나 시기를 고

려하는 중요성에 따라 0~1의 정규화된 값을 사용하여 대량의 데이터에서 정해진 임계값 이상의 빈발 패턴 항목집합을 찾아내는 것이다. 항목별 다른 가중치를 고려해야 하기 때문에, 예를 들어 패턴 ab의 가중치는 (a의 가중치 + b의 가중치)/2가 되고 여기에 ab패턴의 빈도수를 곱하여 해당항목의 가중치 지지도를 구한다. 기존의 가중치 빈발패턴 마이닝 연구[4,7]에서는 빈발패턴 마이닝에서 적용하는 Anti-monotone성질을 가중치 빈도수에 적용할 수 없는 문제점을 개선하기 위해 전역적 최대 가중치를 설정하여 해결하였다. 그리고 [11]에서는 불필요한 항목을 제거하기 위해 전역적 최대 가중치인 GMAXW를 이용하고, 빈발 가능성이 없는 후보 항목 생성을 줄이기 위해 지역적 최대 가중치인 LMAXW를 이용하였다.

온톨로지는 합의된 지식을 표현하고, 일반화하는 것은 가능하지만 사용자들에게 공통적으로 적용되므로 개별적이며, 동적으로 변하는 정보를 반영하는 것에 한계가 있다. 사용자에게 적합한 서비스를 제공하기 위해서는 시간의 흐름에 따라 변화할 수 있는 사용자의 정보 선호 패턴, 사용이력 등을 반영할 수 있는 방법이 필요하다. 그리고 지속적으로 입력되는 스트림 데이터에 대한 마이닝[12,13]에도 적용가능해야 한다. 따라서 본 논문에서는 사용자의 서비스에 대한 관심의 변화를 동적가중치에 반영하는 빈발 서비스 패턴 탐사 방법을 제안한다. 사용자에게 적절한 서비스 정보를 제공하기 위한 기본 연구로써 서비스 정보 온톨로지를 기반으로 마이닝을 수행한다. 제안 방법은 서비스의 종류를 관심있는 도메인으로 세분화하여 사용자의 선호 서비스를 온톨로지로 규칙으로 저장하고, 서비스 정보는 시간과 공간, 사용자의 프로파일 등을 고려한 서비스 빈발 패턴을 발견한다.

3. 동적 가중치 패턴 탐사

컨텍스트 정보를 인지하기 위한 기초가 되는 것은 사용자 주변에 존재하는 환경과 객체들을 컨텍스트 온톨로지로 모델링하고 이를 바탕으로 서비스를 제공하는 것이다. 시간은 공간과 함께 존재한다. 사람의 행동도 시간과 공간이 항상 결부되어 있다. 시공간 이동 패턴은 이동하는 객체의 위치 패턴으로 고객의 위치 특성에 따라 개인화되고 알맞은 컨텐츠나 서비스 제공을 가능하게 하는 시공간 규칙이다[14]. 그러므로 시간과 공간을 함께 고려하는 온톨로지를 통해 사용자의 행동에 대한 서비스 제공이 필요하다. 즉, 시공간 관계는 시공간 객체와 관련된 사건들 간의 인과 관계(casual relationship)를 탐사하는 데 매우 중요한 의미를 가지므로 시공간 특성을 함께 고려하는 온톨로지를 설계한다.

OWL은 class와 property를 표현하는 다양한 어휘들을 제공한다. context entity들 간의 관계와 data들은 property로 표현된다. Entity들 간의 관계는 property로 표현한다. (그림 1)은 시공간 정보를 고려하는 서비스 온톨로지를 표현한 OWL의 일부이다.

(그림 1) 서비스 온토로지의 OWL 표현

```
<owl:Class rdf:ID="Service_Ontology"/>
  <owl:Class rdf:ID="Service">
    <rdfs:subClassOf
          rdf:resource="#Service_Ontology"/>
  </owl:Class>
  <owl:ObjectProperty rdf:ID="ProvidedIn">
    <rdfs:domain rdf:resource="Service">
    <rdfs:range rdf:resource="Location">
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="ProvidedAt">
    <rdfs:domain rdf:resource="Service">
    <rdfs:range rdf:resource="Time">
  </owl:ObjectProperty>
  <owl:DataProperty rdf:ID="Service_ID">
    <rdfs:domain rdf:resource="Service">
    <rdfs:range rdf:resource="xsd:string">
  </owl:DataProperty > ...
  <owl:Class rdf:ID="Guide">
    <rdfs:subClassOf rdf:resource="#Service"/>
  </owl:Class> ...
  <owl:Class rdf:ID="Traffic">
    <rdfs:subClassOf rdf:resource="#Guide"/>
  </owl:Class>
```

(Figure 1) Service Ontology by OWL

사용자를 위한 서비스는 시간과 공간 정보가 연관된 서비스 규칙으로 구성된다. 시간 정보는 일상적인 생활패턴으로 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 등, 추천서비스는 상품 추천 및 숍 추천 등, 예약서비스는 호텔예약 및 운송수단예약 등으로 세분화하였다. 서비스 빈발 패턴의 의미는 해당시간과 위치에서 해당 서비스가 자주 이용된다는 것을 말한다. 예를 들어, 빈발 서비스 패턴 (Sr7, Sp3, St2)은 해당 사용자별 특정 시간과 장소에서 자주 사용되는 서비스를 의미하며, (Sr7, Sp3), (Sr7, St2), (Sp3, St2)도 빈발한 서비스 패턴에 포함된다. 이러한 서비스 규칙은 기준이 되는 특정 시간정보 또는 위치정보에 대한 서비스 계층내에서 시간별, 위치별, 연령별과 같은 관심있는 서비스 연관규칙도 발견할 수 있는 기반이 된다.

스트림 데이터의 마이닝 수행에서 배치(batch)는 일정한 수의 트랜잭션으로 구성되며, 시간의 흐름에 따라 각 서비스 항목의 가중치는 동적으로 변화하는 것으로 가정한다. 이는 실시간으로 검색 순위가 변경되는 것과도 같은 의미로 관심있는 서비스의 중요도를 가중치로 표현한다. 가중치는 각 서비스 항목에 대한 패턴의 가중치로써 다음과 같은 식으로 계산한다.

$$\text{Pwgt(I)} = \frac{\sum_{i=1}^{length(I)} weight(I_i)}{length(I)} \qquad (1)$$

여기서 Pwgt(I)는 항목으로 구성되는 패턴의 가중치를, length(I)는 패턴을 구성하는 항목의 길이, weight(I_i)는 각 항목의 가중치를 의미한다. 즉, Pwgt(I)는 패턴을 구성하는 각 항목의 가중치 합을 패턴의 길이로 나눈 값이 된다. 예를 들어 패턴 ab, 즉 a와 b가 함께 발생하는 패턴의 Iwgt(ab)는 a:0.6, b:0.8 일 때(0.6+0.8)/2=0.7이다. 패턴의 가중치 지지도는 패턴의 가중치에 패턴의 빈도를 곱한 값이다.

이 논문에서는 [11]의 알고리즘을 변형하여 적용한다. 기존 알고리즘과의 차이점은 SPrefix_tree를 생성하기 전에 각 배치별로 먼저 전처리 과정을 통해 빈발 가능성이 없는 항목들은 제거하고 트리를 생성한다. 이는 트리 사이즈가 커지는 것을 미리 방지할 수 있고, 이것은 마이닝 수행 과정의 성능 향상에도 영향을 미친다. 동적으로 변하는 가중치를 적용하는 마이닝에서는 임의의 패턴이 빈발하여도 패턴의 부분 패턴이 반드시 빈발하지 않다. 즉 Anti-monotone 성질을 충족시키지 못한다. 그러므로 이러한 특성을 고려하여 빈발 후보 항목을 줄이는 방법이 중요하다.

전처리 과정에서 빈발 가능성이 없는 개별항목을 가지치기하기 위해서는 개별항목의 가중치만을 고려하여 선별하기 어렵다. 해당 항목의 가중치가 작아도 길이가 2이상의 패턴항목에 대한 가중치는 예측하기 어렵기 때문이다. 즉, 개별항목의 가중치가 임계치에 만족하지 않아 가지치기하면 길이가 2이상의 빈발항목을 발견하지 못하게 되므로 개별항목에 대한 가지치기를 연기하는 것이다. 이를 위해 기존의 논문[11]에서는 전역적 최대 가중치인 GMAXW(Global Maximum Weight)를 사용하여 하였다. 그러나 GMAXW는 항목의 가중치를 최대의 가중치로 과대계상하여 계산하므로 불필요한 후보항목을 생성하여 실제로 후보항목을 줄이는 데 비효율적이다. 따라서 본 논문에서는 전처리를 위해 개별항목이 발생하는 배치에서의 최대 가중치 Lwgt와 빈도를 이용하여 가지치기한다. Lwgt는 GMAXW보다 작거나 같기 때문에 많은 빈발 후보항목을 줄여주는 효과가 있다. 가지치기 하는 기준이 되는 항목의 IMwgt(Item Maximun weight)은 다음의 식과 같다.

$$\text{IMwgt(I)} = \sum_{i=1}^{n} Lwgt(B_i) \times Freq_i(I) \qquad (2)$$

여기서 Lwgt(Local weight)는 항목 I가 발생하는 배치(B)의 항목 중 가장 큰 가중치를 의미하고, n은 배치의 수이고 $Freq_i$(I)는 배치 B_i에서의 항목 발생빈도를 의미한다. 가지치기 기준은 IMwgt(I) < min_wsup이면 제거되어 트리 생성

에서 제외된다. 개별 항목은 다른 항목들과 함께 발생하여 패턴을 구성하는데, 생성가능한 모든 패턴의 가중치는 IMwgt보다는 작거나 같은 값이 나온다는 것을 이용하여 가지치기하는 것이다. 이와 더불어 마이닝 하는 과정에서 빈발항목이 될 가능성을 판단하기 위하여 2차 가지치기하는 과정에서는 개별항목과 연관되어 발생하는 항목중에서 가장 큰 가중치를 가지고 판단하는데 이를 RMwgt(Related Maximum weight)라 하고 후보항목 가능성 여부를 결정한다. 예를 들어 항목의 가중치가 a:0.6, b:0.8, c:0.5일 경우 개별항목 a와 함께 발생하는 항목으로 b, c가 있을 때 가중치가 큰 b의 0.8이 RMwgt가 되고 RMwgt에 빈도를 곱하여 임계치와 비교하고 가지치기 한다.

서비스 온톨로지는 정보의 종류를 포함하고 있고, 마이닝 과정을 설명하기 위해 하위 계층의 상세 서비스를 단순하게 a, b, c ... 등으로 표현하여 마이닝 과정을 설명한다.

전체 온톨로지 설계에서 일부의 구조만을 가지고 알고리즘의 예를 설명한다. 마이닝을 위한 SPrefix-tree는 순서화된 트리이고, 항목의 id, 빈도, 가중치 정보를 헤더 테이블에 저장한다. 일정한 수의 트랜잭션은 배치집합으로 구분되고, 각 트랜잭션은 알파벳 순서의 항목에 따라 차례로 트리의 각 노드에 빈도와 함께 삽입되며 같은 항목이 삽입할 때마다 빈도가 증가한다. 그러므로 트리의 루트부터 단말노드까지의 경로는 각각의 트랜잭션을 구성하는 항목을 나타내고 공통의 구조가 많으면 항목의 빈도는 증가한다.

<표 1> 항목 가중치

Batch	T-id	Items	Weight	
B_1	T_1 T_2 T_3	a, b b, c, d a, d	a: 0.8 b: 0.6 c: 0.2 d: 0.3 e: 0.1 f: 0.2	Lwgt: 0.8
B_2	T_4 T_5 T_6	a, b, c b, c, e c, d, f	a: 0.4 b: 0.7 c: 0.7 d: 0.3 e: 0.4 f: 0.5	Lwgt: 0.7
B_3	T_7 T_8 T_9	c, e a, f c, d, e	a: 0.2 b: 0.3 c: 0.7 d: 0.5 e: 0.3 f:0.6	Lwgt: 0.7

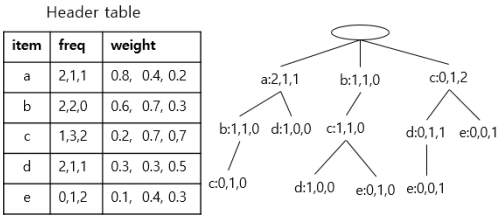
<Table 1> item weight

온톨로지에서 가장 하위레벨의 항목들에 대한 항목레벨을 단순화하여, 날씨:a 교통:b, 숙박:c ,d:공연, 맛집:e, 관광지:f로 표현하고 이들에 대한 각 배치별 트랜잭션에 포함된 항목들에 대한 동적 가중치를 나타낸 것이 <표 1>이다.

SPrefix_tree의 구성방법은 연관규칙에 일반적으로 사용되는 FP트리의 구조와 같고, 트리를 구성하는 자체가 많은 비용이 들기 때문에 트리의 노드를 줄일 수 있도록 전처리를 통해 제외시킬 항목을 가지치기하여 트리를 구성한다.

<표 1>의 예제를 이용하여 각 항목에 대한 IMwgt(I)를 계산하면 a:0.8*2+0.7*1+0.7*1=3, b:0.8*2+0.7*2=3, c:0.8*1+0.7*3+0.7*2=4.3, d:0.8*2+0.7*1+0.7*1=3, e:0.7*1+0.7*2=2.1, f:0.7*1+0.7*1=1.4이다. 마이닝하기 위한 최소 가중치 지지도 min_sup을 1.5라 하면 임계치를 만족하지 못하는 f항목은 가지치기되어 트리 생성에서 제외된다. (그림 2)은 Prefix_tree와 가지치기하여 트리를 생성하는 SPrefix_tree의 예를 보여준다. 트리를 생성하기 전에 빈발 가능성이 없는 항목을 미리 제거하여 트리를 생성하는 비용을 줄일 수 있다.

(그림 2) Batch B1, B2, B3의 SPrefix_tree

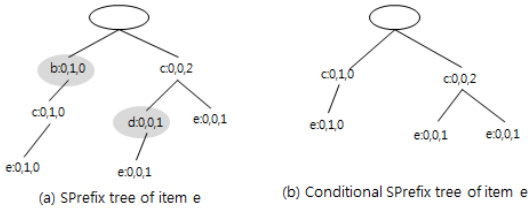


(Figure 2) SPrefix_tree of Batch B1, B2, B3

빈발 구조를 발견하기 위해서는 트리 구조에서 단말노드에 해당하는 항목부터 시작하여 상향노드로 진행하여 각 항목에 대한 빈발 구조패턴을 발견한다. SPrefix_tree에서 연관 발생항목의 최대 가중치인 RMwgt에 빈발도를 곱하여 min_sup를 만족하지 않는 항목은 제거하고 임계치를 만족하는 항목들에 대해서만 조건부 트리를 구성한다. 여기서 RMwgt는 각 기준이 되는 항목들과 함께 발생하는 항목들의 가중치 중에서 가장 큰 값을 의미한다. 그리고 조건부 트리에서 항목들의 실제 가중치를 적용하여 빈발 구

조를 발견한다. (그림 3)의 (a)는 항목 e에 대한 Sprefix_tree이고, 함께 발생하는 연관 항목 b, c, d에서 RMwgt는 0.7이므로 b, d는 각각 0.7*1=0.7이고, c는 0.7*3=2.1이므로 min_sup를 만족하지 않는 b, d는 (그림 3)의 (a)와 같이 삭제되고, 후보항목 e에 대한 조건부 트리는 (그림 3)의 (b)와 같다. 이와 같은 방법으로 다른 항목들에 대해서도 같은 방법으로 적용하여 후보항목를 생성하고 각 항목의 실제 가중치를 적용하여 빈발 항목 여부를 판별한다.

(그림 3) 항목 e의 조건부 SPrefix tree



(Figure 3) Conditional tree of item e

제안하는 마이닝 알고리즘은 다음과 같다.

```
Algorithm: Service mining with dynamic weight
Input: batch series of transactions, min_wsup
Output: frequent item set FS
Begin
  //initialize
 for each transaction Ti in batch set
     Sort items of Ti by alphabet (빈발도)
     Insert items into header table HT(item_id,
                  weight, frequency)
 end for
//preprocessing
 for each item i of transactions in batch Bi
     compute IMwgt(i) on item i
     if IMwgt(i) ≥ min_wsup then
        Candidate item set, Cndt_item = Cndt_item
        ∪ {i};
        insert item i into filtered table FT;
 end for;
//SPrefix_tree 생성
 Make SPrefix_tree of items in FT;
//make conditional tree and frequent items
 for each item i in SPrefix_tree
   make SPrefix_tree(i)
    for each associative item j in SPrefix_tree(i)
      compute RMwgt(j);
     if RMwgt * frequency(j) ≥ min_wsup then
       make conditional_tree CT of item i;
        for each item k in CT(i)
          if wgt(k) * frequency(k)  ≥ min_wsup
             insert item into frequent set FS;
        end for;
     else
        delete item j;
  end for;
 end for;
end;
```

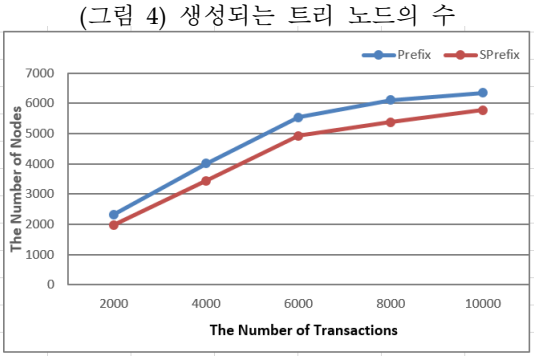
4. 실험

본 논문에서 제안한 상황의 변화에 따라 서비스의 가중치를 동적으로 적용하는 서비스 패턴 마이닝의 효율성을 분석하기 위한 실험을 수행하였다. 데이터는 온톨로지의 단말 노드를 나타내는 서비스 항목들을 입력 데이터로 하여 실험한다. 트랜잭션 수 10,000에 대하여 하나의 배치에 포함되는 트랜잭션의 수는 4로 하고, 전체 데이터 항목에서 각 트랜잭션의 항목은 25-40%의 항목을 포함하는 데이터를 가지고 실험하였다.

첫 번째 실험에서는 트랜잭션의 항목들을 가지고 트리를 생성했을 때의 노드 수를 비교하였다. [11]의 Prefix트리는 모든 트랜잭션의 항목들로 트리를 생성하여 마이닝을 초기화 시킨다. SPrefix트리는 항목에 대한 IMwgt를 가지고 임계치를 만족하지 않는 항목을 전처리 하여, 트리를 생성한다. 생성되는 트리의 노드 수는 트리를 저장하는 메모리 사이즈 및 마이닝의 성능에 영향을 미친다. (그림 4)는 입력되는 트랜잭션 수의 증가에 따라 생성되는 노드 수를 비교한 결과를 보여준다. 결과에서 보는 것처럼 SPrefix는 IMwgt를 적용한 전처리를 통해 빈발 후보 가능성이 없는 항목들을 미리 제거하여 전체 트리의 노드 수가 Prefix에 의한 트리의 노드 수보다 적다는 것을 보여준다. 그리고 트랜잭션 수가 많아질수록 증가되는 노드 수의 비율은 Prefix와 SPrefix에서 모두 점차 줄어든다. 이는 트랜잭션의 항목이 많아지면 중복되는 항목이 많아져서 새롭게 생성되는 노드 수는 상대적으로 많이 증가하지 않는다는 것을 알 수 있다. 그리고

SPrefix에서도 트랜잭션의 수가 증가하면서 중복되는 항목이 많아져서 가지치기되는 항목의 수도 조금씩 줄어드는 것을 수 있었다.

(그림 4) 생성되는 트리 노드의 수

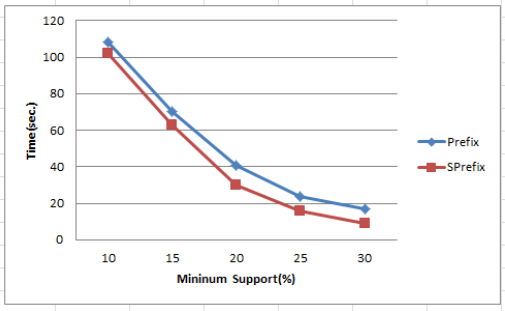


(Figure 4) The number of tree nodes

두 번째 실험에서는 배치에 포함된 트랜잭션 항목의 전체 가중치와 빈도비율(%)을 임계치로 하여 수행 속도를 실험하였다. (그림 5)는 마이닝의 수행 속도 결과를 나타낸다. 그림의 결과에서 마이닝 수행 속도가 Prefix보다 SPrefix가 더 좋은 성능을 보여준다. 이 결과는 첫 번째 실험의 트리를 구성하는 노드 수의 차이가 마이닝의 수행에 영향을 미친다는 것을 보여준다. SPrefix에서는 후보가능성이 없는 항목을 미리 제거하는 전처리 과정이 소요되지만 전체적인 마이닝 시간에는 더 좋은 효과가 있음을 알 수 있었다. 이것은 전체의 트랜잭션 항목에 대해 트리를 구성하는 것보다 전처리 과정을 통해 후보 가능한 항목만을 트리로 구성하는 것이 마이닝 수행에 더 효율적이라는 것을 의미한다. 임계치가 적을 때는 가지치기 되는 항목이 적어 수행시간의 차이가 작지만 임계치가 커질수록 가지치기되는 항목이 증가하여 수행시간이 줄어든다는 것을 알 수 있었다. 그리고 최소 지지도가 커질수록 임계치를 만족하는 항목이 줄어들어 마이닝 수행시간은 점차 줄어든다.

실험결과를 요약하면 제안방법은 항목에 대한 가중치가 일정하지 않기 때문에 후보항목을 되도록 적게 생성하기 위하여 트리를 생성하기 전에 Lwgt를 이용하여 전처리하고 마이닝 과정에서 연관된 항목의 가중치 RMwgt를 이용하여 가지치기하는 방법이 마이닝 수행에 효율적이라는 것을 알 수 있었다. 그리고 제안한 방법은 지속적으로 데이터가 입력되는 스트림 데이터에도 적용할 수 있다는 특징이 있다. 새로운 일련의 배치가 입력되면 가장 오래된 배치의 트랜잭션 항목들의 빈도, 가중치를 포함한 정보를 일괄 제거하고 새로운 입력 항목들의 정보를 포함하여 마이닝을 수행한다.

(그림 5) 마이닝 수행시간



(Figure 5) Running time

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] J. Hipp, U. Guntzer, C. Nakhaeizadeh, “Algorithms for Association Rule Mining - A General Survey and Comparison,” SIGKDD Exploration, 2(1), pp.58-64, 2000

[2] .Agrawal, T. Imielinski and A. Swami, "Mining Asso ciation Rules Between Sets of Items in Large Datab

ase," Proc. of The 12th ACM SIGMOD Int. Conf. on Management of Data, 1993

[3] U. Yun, "Efficient Mining of Weighted Interesting Patterns with a Strong Weight and/or Support Affinity," Information Sciences, Vol. 177, pp.3477-3499, 2007

[4] C. F. Ahmed, S. K. Tanbeer, B. S. Jeong, Y. K Lee, "Mining Weighted Frequent Patterns in Incremental Databases," Proc. of the Pacific Rim, Int. Conf. on Artificial Intelligence, 2008

[5] F. Tao, "Weighted Association Rule Mining using Weighted Support and Significant Framework," Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2003

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

[7] 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

[8] S. Lo, "Binary Prediction based on Weighted Sequential Mining Method," Proc. of the Int'l Conf. on Web Intelligence, pp.755-761, 2005

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

[10] U. Yun, "WIS: Weighted interesting sequential pattern mining with a similar level of support and/or weight", ETRI journal 2007, vol. 29, no.3, pp. 336-352, 2007

[11] B. S. Jeong, A. Farhan: Efficient Dynamic Weighted Frequent Pattern Mining by using a Prefix-tree. The KIPS Transactions, Vol.17, 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