Regular paper http://dx.doi.org/10.13067/JKIECS.2014.9.3.359

데이터의 의미적 정보를 공정하게 반영한 인터트랜잭션들에 대한 연관규칙 탐사

정희택*

Association rule mining for intertransactions with considering fairly data semantics

Hyi-Thaek Ceong*

요 약

최근에는 트랜잭션들 사이의 문맥을 반영하기 위해, 단위 트랜잭션들 사이의 관계를 반영한 확장 트랜잭션을 생성하고 이를 대상으로 인터트랜잭션들에 대한 연관 규칙 탐사방안이 연구되었다. 본 연구에서는 기존 인터트랜잭션들에 대한 연관규칙 탐사 기법에 존재하는 두 가지 문제를 제시하였고 이를 해결하기 위한 방안을 제안하였다. 첫째, 인접한 트랜잭션들 상에 존재하는 데이터의 의미적 변화 정보를 반영하기 위한 방안을 제안했다. 둘째, 트랜잭션을 인터트랜잭션으로 변환하는 과정에서 발생하는 불공정 고려를 해결하기 위한 방안을 제안했다. 이를 통해 기존 연구보다 의미 있는 규칙을 생성할 수 있다. 이를 해양 환경 데이터를 기반으로 실험하여 제시한다.

ABSTRACT

Recently, to reflect the context between transactions, the intertransaction association rule mining has been study. In this study, we present two problems that is within intertransaction association rule mining method and suggest the methods to solve this problems. First, we suggest an algorithm to reflect changes on data between transactions. Second, we propose the method to solve the unfairly considered frequency of data when intertransactions is generate with transactions. We make more meaningful rules than previous researches. We present the experiment result with measured data from the marine environment.

키워드

연관규칙 탐사, 인터트랜잭션, 더미 트랜잭션, 데이터 변화 모델링

Association Rule Mining, Intertransaction, Dummy Transaction, Data Change Modeling

Ⅰ. 서 론

대용량 데이터베이스로부터 전략적 의사결정을 하기 위한 정보의 효과적은 추출은 필수적인 요소가 되었다[1]. 이러한 목적에 부합하기 방안 중에 데이터들로 구성된 트랜잭션들로부터 정보를 발견하기 위해 연관 규칙 탐사 방안이 연구되어 왔다[2-4]. 연관 규칙 탐사 방안은 트랜잭션들로부터 빈발한 데이터를 찾고 이를 기반으로 일정 임계값 이상의 패턴을 발견하기 방안이다[5-7]. 고려 대상 데이터의 특징이 일반 기호적 데이터냐 수치적 데이터냐의 특성을 반영하여 연관 규칙 탐사 방안들이 제안되어 왔다[8-9].

* 교신저자(corresponding author) : 전남대학교 멀티미디어전공(htceong@chonnam.ac.kr)
접수일자 : 2013. 12. 20 심사(수정)일자 : 2014. 02. 20 게재확정일자 : 2014. 03. 10

연관 규칙 탐사 대상은 인트라트랜잭션(intra transaction)과 인터트랜잭션(intertransaction)으로 구분할 수 있다[10-11]. 인트라트랜잭션은 데이터의 묶음인 각각의 트랜잭션들을 대상으로 한다. 일련의 사건들의 발생을 의미하고 있는 트랜잭션들에서 그 사건들의 빈발정도를 기반으로 연관규칙을 발견하는 방안이다. 이러한 접근법은 Apriori를 기본으로 하는 방안이다[11-12]. 이러한 접근법은 트랜잭션이 발생한 문맥(즉, 시간, 장소, 순서)을 고려하지 않다. 이와 달리 인터트랜잭션은 문맥을 반영하기 위해, 단위 트랜잭션들 사이의 관계를 반영한 확장 트랜잭션(즉, 인터트랜잭션)을 생성하고 이를 대상으로 연관 규칙 탐사를 수행한다. 즉, 인접한 단위 트랜잭션들 사이의 순서 관계를 반영한 변환을 통해, 인터트랜잭션 생성한다. 인터트랜잭션에 의한 규칙의 생성은 시간적 정보와 같은 부가적 정보를 생성하게 한다. 해양 환경 데이터를 대상으로 예를 들면, 인트라트랜잭션에 대한 연관 규칙 탐사는 “염도가 29이고 DO가 7이면 수온이 22이다.”와 같은 규칙을 생성할 수 있으나, 인터트랜잭션에 대한 연관 규칙 탐사는 “염도가 29이고 DO가 7이면 *6일 후* 수온이 22이다.”와 같은 규칙을 생성할 수 있다. 후자가 전자보다 많은 정보를 제공할 수 있다.

인터트랜잭션을 대상으로 한 기존 연관 규칙 탐사 방법은, 두 가지 문제를 간과하고 있다. 첫째는 인터트랜잭션을 구축하는 과정에서, 데이터의 변화에 대한 의미적 정보를 반영하지 않는다. 인터트랜잭션을 구축하는 과정에서 단위 트랜잭션들 사이의 순서관계를 반영할 뿐, 각 단위 트랜잭션을 구성하고 있는 데이터가 인접한 데이터와의 관계를 반영하지 않고 있다. 구체적인 예는 2장에서 제시한다. 둘째는 단위 트랜잭션으로부터 인터트랜잭션을 구축하는 과정에서 불공정한 빈도와 관계의 고려가 존재한다. 단위 트랜잭션들에 대해 설정된 슬라이딩 윈도우 크기 단위로 인터트랜잭션을 생성할 때, 윈도우의 시작위치에 따라 단위 트랜잭션의 항목들의 순서 관계 및 빈도를 부분적으로 반영한다. 구체적인 예는 2장에서 제시 한다.

본 연구에서는 앞서 제시한 이러한 문제를 해결할 수 있는, 인터트랜잭션에 대한 연관규칙 탐사 기법을 제안한다. 첫째, 인접한 트랜잭션들 상에 존재하는 데이터의 의미적 변화 정보를 반영하기 위한 방안을 제안한다. 둘째, 트랜잭션을 인터트랜잭션으로 변환하는 과정에서 발생하는 불공정 고려를 해결하기 위한 방안을 제공한다. 이를 통해, 기존 연구에서는 “염도가 29이고 DO가 7이면 6일 후 수온이 22이다.”란 형태의 정보를 탐사할 수 있으나, 본 연구를 통해 ““염도가 29이고 DO가 7을 *이틀 유지하면* 6일 후 수온이 22이다.”란 정보를 탐사할 수 있다.

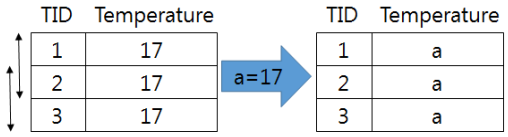
본 논문의 구성으로 2장에서는 인터트랜잭션을 대상으로 한 기존 연구를 분석하고 문제들을 제시한다. 3장에서는 데이터의 의미적 정보를 공정하게 반연한 연관 규칙 탐사방안을 제시한다. 4장은 앞서 제안한 탐사 방안을 해양 환경 데이터를 대상으로 수행하고 그 의미를 제시한다. 마지막으로 5장에서 결론 및 향후 연구를 기술한다.

II. 관련 연구

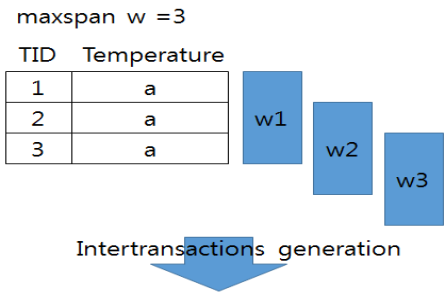
인터트랜잭션에 대한 기존 연구의 분석은 데이터의 의미적 정보를 반영하기 위한 관점과 데이터 항목의 불공정 고려 관점에서 분석한다. 분석함에 있어 인터트랜잭션에 대한 정의 및 알고리즘의 기본은, 인터트랜잭션을 제안한 [10-12]의 연구들을 따랐다. 이는 관련 개념들의 정의를 재 정의하고 기술함으로써 발생하는 공간의 낭비를 최소화하고 본 연구에서 제시하고자 하는 문제에 집중하기 위함이다.

인터트랜잭션에서 데이터의 의미적 정보의 반영은 단위 트랜잭션들 사이의 순서 관계만을 반영할 뿐, 단위 트랜잭션을 구성하는 데이터들 사이의 의미적 관계를 반영하지 못한다. 예를 들면, 그림 1과 같이, 해양 환경의 수온 데이터 집합을 고려할 때, 수치 데이터를 기호로 변환한 후 인터트랜잭션을 생성할 수 있다. 여기서 최대 슬라이딩 윈도우의 크기(maxspan)을 3으로 가정하였다. (a)는 단위 트랜잭션의 데이타를 기호로 변환한 것이며 (b)는 인터트랜잭션 생성을 제시한 것이다. 생성된 인터트랜잭션에서 임계 지지도를 50%라 가정할 때 빈발 항목은 ‘a(0), a(1)’이다. 이는 의미적으로 수온 17도가 첫 번째 존재하는 경우와 두 번째 존재하는 경우가 빈발했음을 의미한다. 이는 데이터 사이의 변화 정보를 반영하지 못한다. 즉, 17도가 연속하여 두 트랜잭션에서 반복된다는 정보를 고려하지 않고 있다. 이른 그림 1의 (a)에서 두 개의

화살표로 표현된 부분에서 확인할 수 있다.



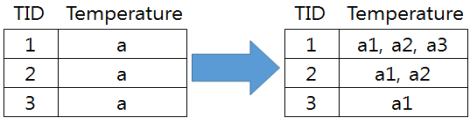
(a) 수치 데이터의 변환
(a) Numerical data transform



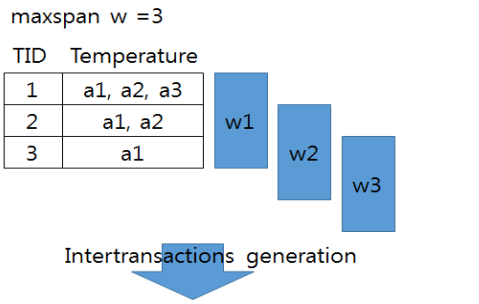
(b) 인터트랜잭션 생성
(b) Intertransaction generation

그림 1. 인터트랜잭션 생성
Fig. 1 Intertransaction generation

단위 트랜잭션을 구성하는 데이터 사이의 변화를 고려함으로써, 단위 트랜잭션들 사이의 순서 관계뿐만 아니라 데이터 사이의 의미를 반영할 수 있다. 이를 예로 들면 그림 2와 같다. 먼저 (a)에서 각 데이터의 변화를 모델 하였다. 즉 'a2'은 상태 'a'가 두 트랜잭션에 걸쳐 존재함을 의미한다. 예에서는 수온 17도가 삼일 동안 변화 없음을 의미하는 것은, TID 1의 'a3'로 모델링 되었다. 데이터의 변화를 반영한 모델일 반영한 단위 트랜잭션들에 대한 인터트랜잭션 생성은 (b)와 같다. 그 결과 생성된 인터트랜잭션들에서 빈발항목은 'a1(0), a2(0), a1(1)'이다. 'a1(0)'과 'a1(1)'은 기존 연구와 동일하게 수온 17도가 첫 번째 존재하는 경우와 두 번째 존재하는 경우가 빈발했음을 의미한다. 다만 시간적 관점에서 본다면 1일 동안 유지되었음을 강조하고 있다. 그러나 데이터들 사이의 변화를 모델링함으로써, 새로 발견된 'a2(0)'는 수온 17도가 이틀 연속하여 존재함을 발견할 수 있게 한다. 단위 트랜잭션의 경계를 허무는 인터트랜잭션 관점에서 볼 때, 단위 트랜잭션을 구성하는 데이터의 경계를 허물고 그들 사이의 의미 있는 변화를 고려한 항목의 생성은 필요하다. 각 데이터의 변화를 반영한 구체적인 방안은 3장에서 제안한다.



(a) 데이터 변화를 기반으로 한 변환
(a) Transform based data variation

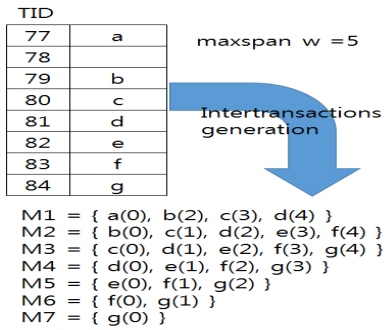


(b) 확장된 항목에 대한 인터트랜잭션 생성
(b) Intertrasnaction generation for extended items

그림 2. 확장된 항목에 대한 인터트랜잭션 생성
Fig. 2 Intertrasnaction generation for extended items

다음으로 인터트랜잭션을 구성할 때, 윈도우 단위보다 앞선 순서의 단위 트랜잭션의 데이터들 사이의 빈도 및 순서 관계가 불공정하게 고려된다. 이를 예로 들면 그림 3과 같다. 먼저 단위 트랜잭션들로부터 인터트랜잭션의 생성은 (a)에서 보이고 있다. 생성된 인터트랜잭션들을 고려할 때, 데이터 자체의 출현 빈도를 고려하면 (b)와 같은 결과를 확인 할 수 있다. 즉, 항목 'a'는 1회 고려되었음에 비해 항목 'f'나 'g'는 5회 고려되었다. 또한 항목 'a'와 'b'의 관계를 고려할 때, 'a(0),b(2)', 'a(1),b(3)', 'a(2),b(4)'의 모두 패턴을 고려해야 하나, 기존 연구에서는 그림에서 확인할 수 있듯이 'a(0),b(2)'만을 고려하고 있다. 이는 슬라이딩 윈도의 시각을 항목이 있는 트랜잭션부터 시작하는 인

터트랜잭션 생성 방법에 기인한다. 이를 해결하기 위한 방안은 3장에서 제시한다.



(a) 인터트랜잭션 생성
(a) Intertransaction generation

Symbol	Considered frequency
a	1
b	2
c	3
d	4
e	4
f	5
g	5

(b) 각 기호에 대한 고려 횟수
(b) Considered counters of each symbol

그림 3. 불공정하게 고려되는 기호들
Fig. 3 Unfairly considered symbols

III. 의미적 정보를 공정하게 반영한 인터트랜잭션을 활용한 연관규칙 탐사

본 장에서는 앞서 제시한 문제점들을 해결할 수 있는 방안을 제안한다. 첫째, 단위 트랜잭션의 순서관계를 반영하면서로 데이터 차원의 변화정보를 모델링하기 위한 방안을 제안한다. 다음으로는 인터트랜잭션을 생성하는 과정에서 발생하는 항목의 고려 횟수의 불공정성을 해결하기 위한 방안을 제안한다. 이는 불공정하기 고려되는 항목은 그 결과로 생성되는 규칙도 원래 데이터가 갖는 정보를 필요 충분히 반영하지 못하는 문제를 야기하기 때문이다.

3.1 데이터의 변화를 반영한 인터트랜잭션 생성

여기서는 단위 트랜잭션들을 구성하는 데이터 상이의 변화를 모델링하기 위한 방안을 제안한다. 동일한 속성에 속하는 데이터들의 변화는 그 형태에 따라 구분할 수 있다. 비수치 데이터와 수치 데이터로 크게 구분할 때, 수치 데이터의 변화가 비수치 데이터의 변화보다 다양한 의미와 정보를 대표할 수 있다. 여기서는 수치 데이터의 변화를 모델링 하는 방안을 중심으로 기술한다. 비수치 데이터인 경우는 수치 데이터인 경우보다 변화의 종류가 적기 때문에 대표하는 기호로 변환 후 간단하게 모델링 할 수 있다.

수치 형태의 데이터들의 변화를 모델링하기 위해, 데이터의 변화는 3가지 형태로 구분할 수 있다. 한 데이터의 값이 다음 값(즉, 다음 순서에 있는 트랜잭션에 있는 동일 속성의 값)에서 커지거나, 작아지거나, 또는 동일한 값의 변화를 갖는다. 커지거나 작아지는 값의 변화는 한 단위 트랜잭션들 사이의 변화를 의미하는 것으로, 현재 데이터에 이 변화를 모델링한다. 이를 다음과 같이 데이터 값의 변화와 데이터 값의 유지라 정의한다. 일반성에 위배됨이 없이 트랜잭션을 구성하는 스키마가 n개의 속성으로 구성된 $S=<A1,A2,...,An>$을 가정할 때, 임의의 두 트랜잭션을 $T_i=<v_i^{A1},v_i^{A2},...,v_i^{An}>$, $T_j=<v_j^{A1},v_j^{A2},...,v_j^{An}>$이라고 하자. 여기서 V_i^{Am}은 트랜잭션 T_i에서 속성 Am의 값을 의미 한다($1 \leq m \leq n$). 이를 기반으로 정의하면 다음과 같다.

정의 1. 데이터 값의 변화

$j=i+1$일 때, $v_i^{Am} \neq v_j^{Am}$이면, 인접한 두 트랜잭션 T_i와 T_j은 속성 Am에서 데이터 값의 변화라고 정의한다.

정의 2. 데이터 값의 유지

$j=i+1$일 때, $v_i^{Am}=v_j^{Am}$이면, 인접한 두 트랜잭션 T_i와 T_j은 속성 Am에서 데이터 값의 유지라고 정의한다.

데이터 값의 변화는 인접한 데이터 값들의 변화만이 존재하지만, 데이터 값의 유지는 지속적으로 발생할 수 있다. 즉, 인접한 두 데이터들이 이행성(transitive) 특징을 갖는다. 예를 들면 T_i, T_{i+1}, T_{i+2}에서 $v_i^{Am}=v_{i+1}^{Am}=v_{i+2}^{Am}$에 의해 데이터 값이 유지가 지속될 수 있다. 이를 반영한 모델링이 이루어져야 한

다.

데이터 값의 변화와 이행성을 갖는 데이터 값의 유지를 모델링하기 위해, 다음과 같은 알고리즘을 제안한다. 알고리즘의 구성은 먼저, 각 데이터를 기호로 변경한다. 변경된 각 속성의 데이터에 대한 데이터 값의 변화 및 유지와 관련된 정보를 모델링한다. 여기서 데이터 값의 변화도 이행성 특징을 고려할 수 있으나, 이는 동일 속성에 속하는 인접 데이터의 변화가 일정한 패턴으로 변화는 경우를 고려할 수 있다. 그러나 이러한 경우는 일반적이지 않기 때문에 고려대상에서 제외하였다.

```
Symbolic transform based on Data change pattern
Input : Transaction T_i database( 1 ≤ i ≤ m )
Output : Extended transaction database

SD //Symbol database
for k=1 to n
   for I =1 to m
      s_i^Ak = SelectSymbol(v_i^Ak, SD);
      //s_i^Ak is a symbol of v_i^Ak.
      // If there is the symbol of v_i^Ak in SD, it return.
      // Otherwise, it select new symbol.
      ms = ExchangeItemSymbol(s_i^Aks, s_i^Ak+'1');
      //s_i^Ak is replaced as 's_i^Ak1'
      TransformValueBySymbol(v_i^Ak, ms);
   end for
end for

for k=1 to n
   for i = 1 to m
      if (s_i^Ak ≠ s_i+1^Ak) {
         MCI = MakeItem(s_i^Ak, s_i+1^Ak);
         // concatenate(s_i^Ak, s_i+1^Ak)
         AddItemInTransaction(MCI, T_i);
      } else {
         for t = 2 to m - i
            MCI = MakeItemWithReplacement(s_i+t^Ak, t);
            //'s_i^Ak1 is replaced as 's_i^Ak2',s_i^Ak3,
            // or s_i^Akm based on t..
            AddItemInTransaction(MCI, T_i);
            if (s_i^Ak ≠ s_i+t+1^Ak ) break;
         end for
      }
   end for
end for
```

제시한 알고리즘은 먼저, 각 데이터를 기호로 변환하는 블록과 인접한 데이터의 변화를 모델링하는 블록으로 이루어졌다. 제시한 알고리즘의 수행과정을 쉽게 설명하기 위해, 그림 4에 제시한다. 그림 4는 데이터 값의 변화와 데이터 값의 유지를 모델링하는 예이다. 먼저 각 트랜잭션에 있는 데이터 값은 기호로 변환되며, 다음으로 인접한 데이터와의 변화를 반영한 기호를 생성하게 된다. TID 22에서 'y1,y2,y3'의 의미는 '수온이 17도인 상태가 하루, 이틀, 사흘 계속됨'을 나타낸다. TID 25의 'z1, z1y1'은 '수온이 15도인 날이 하루 있었으며 그 다음날을 16도로 변경되었음'을 나타낸다.

TID	Temperature
21	16
22	17
23	17
24	17
25	15
26	16

TID	Temperature
21	x1, x1y1
22	y1,y2,y3
23	y1,y2
24	y1,y1z1
25	z1,z1y1
26	y1

그림 4. 데이터 값의 변화에 대한 변환 예
Fig. 4 Transform example of data values change

한편, 비수치 데이터인 경우는 상태 전이 형태로 모델링 할 수 있다. 예를 들면 '자동차 소유형태'란 속성에 대한 'owner','rent','lease' 중 하나를 갖는다면 각 상태의 변화로 모델링 할 수 있다. 위 예에서 TID 21, 24, 25에서와 동일한 방법으로 그 의미를 반영할 수 있다.

3.2 공정한 인터트랜잭션 생성 방법

인터트랜잭션을 구성할 때, 윈도우 단위보다 앞선 순서의 단위 트랜잭션 데이터들의 빈도가 불공정하게 고려된다. 앞서 그림 3에서 예로 제시한 바와 같이, 임의의 슬라이딩 윈도우 크기 w와 트랜잭션 개수 x에 대해, 데이터 항목은 [min(w,x-(w-1))-1, w]사의 빈도로 고려된다. 생성된 인터트랜잭션 상에 모든 데이터 항목이 동일한 빈도를 갖지 못한다. 또한 슬라이딩 윈도우의 시작이 항목이 있는 트랜잭션으로부터 시작함으로써 항목들 사이의 관계 패턴을 고려하지 못한다. 이를 본 연구에서는 다름과 같이 불공정 고려로 정의한다.

정의 3. 데이터의 불공정 고려 문제

인터트랜잭션을 생성할 때, x개의 트랜잭션들에 대해 윈도우의 크기가 w에서, 단위 트랜잭션에 있는 데

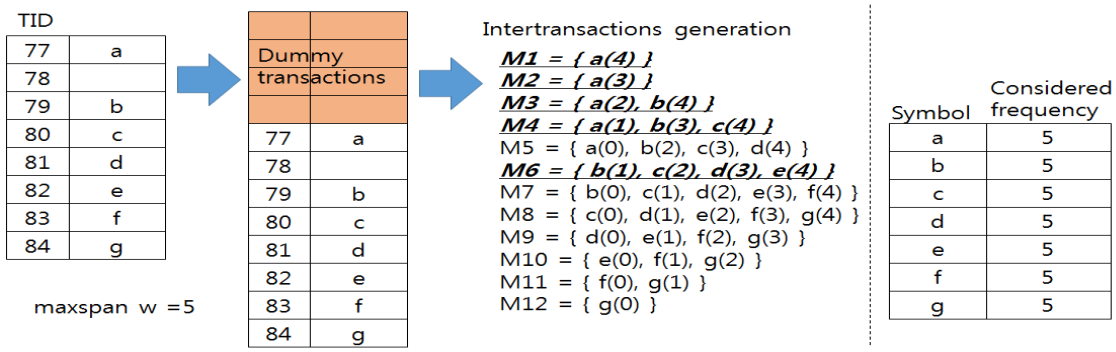


그림 5. 더비 트랜잭션을 고려한 인터트랜잭션 생성
Fig. 5 Intertransaction generation based on considering dummy transactions

이터들은 빈도를 동일하게 반영하지 않고, 존재할 수 있는 모든 순서 관계를 로 고려하지 않는 문제.

이러한 불공정 고려문제를 해결하기 위해서, 본 연구에서는 더미(dummy) 트랜잭션 기반 슬라이딩 윈도우 기법을 제안한다. 기존 인터트랜잭션 생성방안과 달리, 트랜잭션들 사이의 모든 순서 패턴을 반영하기 위해 더미 트랜잭션을 추가하고, 데이터가 없는 트랜잭션을 윈도우 시작 지점으로 허용한다. 이를 통해 앞선 예에서 제시한 데이터의 불공정 고려문제를 해결할 수 있다. 더미 트랜잭션 기반 슬라이딩 윈도우 기법을 활용한 인터트랜잭션 생성은 먼저, 윈도우 크기보다 1 작은 더미 트랜잭션을 제일 위에 있는 단위 단위트랜잭션 앞에 추가한다. 데이터가 없는 더미 트랜잭션부터 인터트랜잭션을 생성한다. 트랜잭션들 사이에 데이터가 없는 트랜잭션에 대해서도 그것부터 시작하는 인터트랜잭션을 생성한다. 단순한 방법으로 데이터의 불공정 문제를 해결할 수 있기 때문에, 예를 통해 그 과정과 의미를 제시한다.

데이터의 불공정 고려 문제를 해결하는 방안의 예는 그림 5와 같다. 윈도우 크기보다 1작은 4개의 더미 트랜잭션들을 추가하고, 인터트랜잭션의 시작을 데이터가 없는 트랜잭션에서 시작을 하용하며 데이터의 불공정 고려 문제를 해결하였다. 새로이 생성된 인터트랜잭션은 M1,M2,M3,M4, M6이다. 각 기호에 대한 고려 횟수도 모두 동일하게 5회이며, 생성된 인터트랜잭션에서 항목 'a'와 'b'사이의 모든 순서 관계 'a(2),b(4)', 'a(1),b(3)', 'a(0),b(2)'를 발견할 수 있다.

3.3 인터트랜잭션을 활용한 연관 규칙 탐사

데이터의 변화를 반영하면서 공정한 인터트랜잭션을 생성하기 위한 과정은, 먼저 수치 데이터에 대해 그 변화를 모델링한다. 이는 앞서 제시한 데이터 값의 변화와 유지 형태를 각각 반영하여 수행한다. 다음으로 더미 트랜잭션을 추가하고 데이터가 없는 트랜잭션도 슬라이딩 윈도우의 시작점으로 허용하는 인터트랜잭션을 생성한다. 생성된 인터트랜잭션에 대해 [11-12]에서 제안한 빈발 항목 발견을 통해 연관규칙을 생성한다.

본 연구에서는 데이터의 변화 및 유지를 모델링하고 더미 트랜잭션을 허용함으로써, 기존 인터트랜잭션 생성방법 보다 많은 데이터 항목 및 인터트랜잭션들을 생성한다. 이는 데이터의 일련의 변화를 모델링하기 위해 필요한 과정이다. 특히 데이터가 없는 트랜잭션과 더미 트랜잭션으로부터 인터트랜잭션의 생성을 허용함으로써, 기존 연구에서 고려하지 못한 데이터의 불공정 고려 문제를 해결할 수 있다. 추가된 항목들과 인터트랜잭션들에 대한 수행 성능 문제는 [13-14] 연구들에서 제안한 방안들을 이용하여 완화할 수 있다.

인터트랜잭션이 단위 트랜잭션의 벽을 해소했다는 관점에서, 기존 연구들은 트랜잭션들의 모든 순서 관계를 고려하지 못하였기 때문에, 본 연구에서 제안한 방안이 변환과정에서 정보의 놓침이 없는 의미 있는 인터트랜잭션 생성 방안이다.

Ⅳ. 실 험

본 장에서는 앞서 제시한 문제와 해결 방안을 반영한, 인터트랜잭션 연관 규칙 탐사를 보이기 위해, 2009년 7월부터 8월까지 YSI 해양 센서를 이용하여 남해안에서 하루에 한번 측정한 데이터를 대상으로 하였다. 이는 계절적인 변화가 가장 많은 시간을 대상으로 하였고 측정된 정보는 수온, 염도, 용존산소, pH이다. pH는 대상기간동안 '8'로 동일한 상태를 유지하기 때문에 제외하였다. 단순화를 위해 측정치의 소수점 아래는 반올림하였다.

데이터에 대한 실험은 기존 인터트랜잭션 연관 규칙 탐사 방법과 본 연구에서 제안한 방안을 대상으로 하였다. 이때 생성된 기호와 인터트랜잭션의 개수는 표 1과 같다. 두 기법에 있어 최대 슬라이딩 윈도우 크기는 7로 하였다.

표 1. 실험 매개변수
Table 1. Parameters for experiments

Parameters \ Method		Original Intertransaction Method	Proposed Method
Symbols	Temperature	8	19
	Salinity	4	10
	DO	6	15
# intertransactions		61	67
maxspan		7	7

본 연구에서 제안한 데이터의 값의 변화와 유지를 모델링하기 위해 수온은 19개, 염도는 10개, 용존산소는 15개의 기호를 사용하였다. 데이터의 불공정 고려 문제를 해결할 수 있는 67개의 인터트랜잭션들을 생성한 후, 1 빈발항목, 2 빈발항목, 3 빈발항목을 생성하였다. 최소지지도(min support)를 15과 20으로 구분하였고 각각에 대해 생성된 빈발항목의 개수는 표 2와 같다. 제안한 기법이 기존 기법보다 많은 1 빈발 및 2 빈발항목을 생성하였다. 이는 데이터의 변화를 모델링하였기 때문이다.

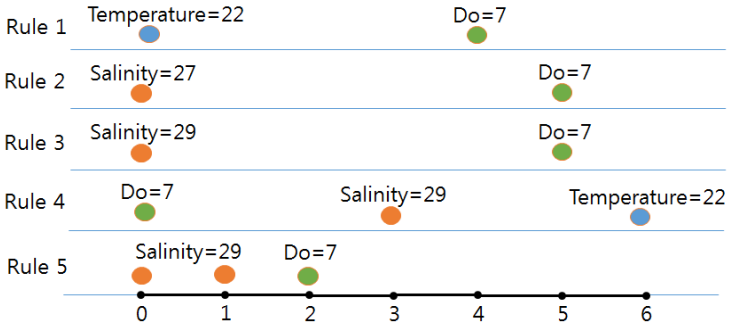
표 2. 빈발 항목 생성
Table 2. Frequent itemsets generation

Original Intertransaction Method			Proposed Method	
15	20	min support	15	20
29	16	Frequent 1-itemsets	133	63
95	3	Frequent 2-itemsets	651	86
18	-	Frequent 3-itemsets	-	-
7	-	Frequent 4-itemsets	-	-
-	-	Frequent 5-itemsets	-	-

기존 인터트랜잭션 기법은 순수 빈발항목만을 고려할 때, 4빈도 항목까지 생성하지만, 본 연구에서 제안한 방안은 2빈도항목까지 생성할 수 있었다. 생성된 빈발항모들에 대해 [11-12] 연구에서 제시하는 규칙 생성 기준에 의해 규칙을 생성하면 그림 6과 같다. 많은 규칙 중 대표적인 규칙을 중심으로 제시하였다. 각 규칙을 시각적으로 확인하기 쉽게 하기 위해, 그림으로 표현하였다. 한편, 본 연구에 의해 생성된 규칙의 유용성과 의미를 설명하기 위해, 최소 지지도 15 주변의 빈도수를 갖는 3빈발 항목과 2빈발항목에 대한 규칙들을 제시하였다. 각각의 기호를 정리하면 'ta'는 수온 21도를, 'tb'는 수온 22도를, 'de'는 용존산소가 7임을, 'sa'는 염도가 27임을, 'sb'는 염도가 28임을, 'sc'는 염도가 29임을 의미한다. 각 기호 뒤에 붙어있는 각 숫자는 해당 상태의 유지기간을 의미한다. 여기서는 하루 한번 측정된 데이터를 사용하였기 때문에 'sc3'은 3일간 염도가 29를 유지하였음을 의미한다. 생성된 규칙 'tb2(0)de1(6)'의 의미는 수온이 22도에서 2일 동안 유지된 후 4일 후에 용존산소가 7임을 의미한다. 규칙 'de1(0)sc1(0)tb2(1)'은 용존산소가 7이고 염도가 29이면 하루 후에 수온이 22도를 2일 동안 유지함을 의미한다.

각 생성된 빈발항목에 대한 규칙들에서 확인할 수 있듯이, 본 연구에서 제안한 방안이 보다 데이터의 변화를 잘 발견할 수 있다. 특히, 기존 인터트랜잭션 기법에 의한 규칙 5는 '염도 29 상태가 이틀 유지되었음'을 알 수 있으나, 이를 위해 규칙에 대한 재처리 과정이 필요하다. 그러나 본 연구에서 제안한 방안은

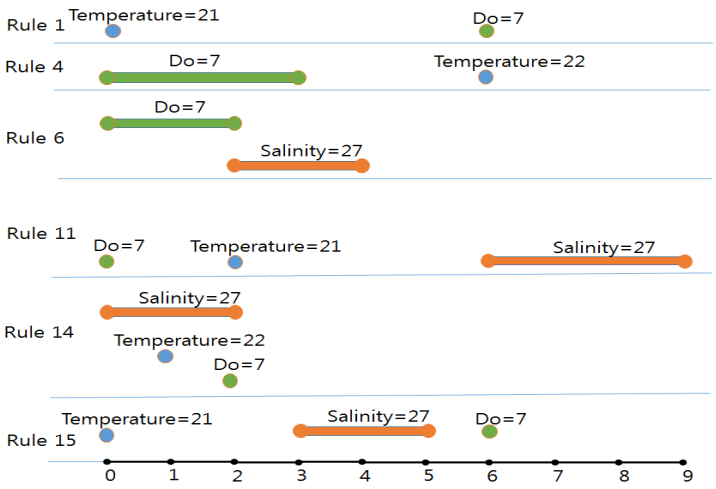
No	rule set	support	confidence
1	tb(0)de(4)	17	85%
2	sa(0)de(5)	16	89%
3	sc(0)de(5)	18	67%
4	de(0)sc(3)tb(6)	10	22%
5	sc(0)de(1)sc(1)	19	70%
6	sc(0)de(3)tb(6)	10	37%



(a) 기존 인터트랜잭션 기법의 규칙들과 표현

(a) Rule set and it's representation of original intertransaction method

No	rule set	support	confidence
1	ta1(0)de1(6)	15	94%
2	tb2(0)de1(6)	16	94%
3	de1(0)tb2(4)	15	30%
4	de3(0)tb1(6)	16	43%
5	de1(0)sb1(6)	15	30%
6	de2(0)sc2(2)	15	37%
7	sa3(0)de1(4)	15	94%
8	sc2(0)de1(5)	17	74%
9	sc3(0)de1(6)	15	79%
10	de1(0)sc1(0)tb1(6)	13	54%
11	de1(0)ta1(2)sa3(6)	11	22%
12	de1(0)sc1(0)tb2(1)	11	22%
13	sa3(0)ta1(3)de1(6)	10	45%
14	sc2(0)tb1(1)de1(2)	12	52%
15	ta1(0)sa2(3)de1(6)	10	63%
16	tb1(0)sa1(3)de2(6)	10	45%



(b) 제안한 기법의 규칙과 표현

(b) Rule set and it's representation of proposed method

그림 6. 결과 규칙 집합과 규칙들의 그래픽 표현

Fig. 6 Result rule set and graphic representation

그림 6(b)에서와 같이 바로 확인할 수 있다.

V. 결 론

데이터들로 구성된 트랜잭션들로부터 정보를 발견하기 위해 연관 규칙 탐사 방안이 연구되어 왔다. 최근에는 트랜잭션들 사이의 문맥을 반영하기 위해, 단위 트랜잭션들 사이의 관계를 반영한 확장 트랜잭션을 생성하고 이를 대상으로 인터트랜잭션들에 대한 연관 규칙 탐사방안이 연구되었다.

본 연구에서는 기존 인터트랜잭션들에 대한 연관규칙 탐사 기법에 존재하나는 두 가지 문제를 제시하였고 이를 해결하기 위한 방안을 제안하였다. 첫째, 인접한 트랜잭션들 상에 존재하는 데이터의 의미적 변화 정보를 반영하기 위한 방안을 제안했다. 둘째, 트랜잭션을 인터트랜잭션으로 변환하는 과정에서 발생하는 불공정 고려를 해결하기 위한 방안을 제안했다. 이를 통해 의미 있는 규칙의 발견을 보이기 위해, 해양환경 데이터를 기반으로 실험하였으며, 그 결과 규

칙을 제시하였다.

본 연구에서는 기존 연구와 달리 특정 데이터 상태의 변화를 발견할 수 있다. 기존 연구에서 '용존산소가 7이고 염도가 29이면 하루 후에 수온이 22도 이다'라는 정보를 발견할 수 있음에 비해, 본 연구에서는 '용존산소가 7이고 염도가 29이면 하루 후에 수온이 22도를 2일 동안 유지 한다'는 정보를 발견할 수 있다.

향후 연구서는 다양한 데이터 집합에 대한 적용을 통해, 데이터 값의 변화를 보다 효과적으로 모델링하는 방안을 연구하고자 한다. 또한 생성된 규칙들을 효과적으로 시각화 하는 방안이 필요하다.

참고 문헌

[1] H. J. Altman, R. B. Kumar, V. H. Mannila, and D. Pregibon, "Emerging scientific applications in data mining," *Communications of the ACM*, vol. 45, no. 8, 2002, pp. 54-58.

[2] A. Tung, L. Hongjun, J. Han, and L. Feng, "Efficient mining of intertransaction association rules," *IEEE Trans. Knowledge and Data Engineering,* vol. 15, no. 1, 2003, pp. 43-56.

[3] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *SIGMOD Rec.* vol. 22, no. 2, 1993, pp. 207-216.

[4] Z. J. Mohammed, "SPADE: An Efficient Algorithm for Mining Frequent Sequences," *Machine Learning Journal*, vol. 42, 2001, pp. 31 - 60.

[5] D.-J. Chai, K. Ban, and E.-K. Kim, "Schema Mapping Method using Frequent Pattern Mining," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 5, no. 1, 2010, pp. 93-101.

[6] S. Kotsiantis and D. Kanellopoulos, "Association rules mining: A recent overview," GESTS Int. *Trans. on Computer Science and Engineering*, vol. 32, no. 1, 2006, pp. 71-82.

[7] B. Nayak and S. Prasad. "A Study of Association Rule Data Mining Approaches and its Current Research Directives on Large Databases," *Int. Review on Computers & Software*, vol. 6, no. 5, 2011.

[8] M. Vannucci and V. Colla, "CollaMeaningful discretization of continuous features for association rules mining by means of a SOM," *In ESANN*, Apr. 2004, pp. 489-494.

[9] K. Taboada, K. Shimada, S. Mabu, K. Hirasawa, and J. Hu, "Association rule mining for continuous attributes using genetic network programming," *In SICE Annual Conf.*, 2007, pp. 2723-2729.

[10] C. Berberidis, L. Angelis, and I. Vlahavas, "Inter-transaction association rules mining for rare events prediction," *In Proc. 3rd Hellenic Conf. on Artificial Intellligence*, 2004.

[11] Y.-P. Huang, L.-J. Kao, and F.-E. Sandnes, "Efficient mining of salinity and temperature association rules from ARGO data," *J. Expert Systems with Applications*, vol. 35, no. 1-2, 2008, pp. 59-68.

[12] V. P. Arunachalam and S. Karthik, "A Novel Approach For Minming Inter-transaction Itemsets," *European Scientific Journal*, vol. 8, no. 14, 2012.

[13] W. Yang, Y. Li, and Y. Xu, "Granule based inter-transaction association rule mining," *Intechopen*, 2008.

[14] D. Bhanu and P. Balasubramanie, "Predictive Modeling of Inter-Transaction Association Rules-A Business Perspective," *IJCSA*, vol. 5, no. 4, 2008, pp. 57-69.

저자 소개

정희택(Hyi-Thaek Ceong)

1992년 2월 전남대학교 전산통계학과 학사

1995년 2월 전남대학교 전산통계학과 석사

1999년 8월 전남대학교 전산통계학과 박사

1999년~현재 : 전남대학교 멀티미디어전공 교수

※ 관심분야 : RFID/USN, 데이터마이닝, 분산처리시스템, 디지털 방송