흥미있는 항목모임의 탐색알고리듬

배철진, 공혜옥

최근 대량의 자료에 대한 분석수법인 자료발굴모형에 대한 연구를 강화하는것은 대단히 중요한 문제로 나선다.

론문에서는 정 및 부의 련관규칙발굴의 첫단계에서 해결해야 할 흥미있는 정 및 부의 항목모임들을 탐색하는 알고리듬을 제기한다. 전통적인 련관규칙발굴알고리듬들은 빈발항목모임들에 토대한 정의 련관규칙들을 추출하며 따라서 먼저 빈발항목모임들 즉 흥미있는 정의 항목모임들만을 탐색하게 된다. 더우기 전통적인 알고리듬들에서 탐색공간을 줄이기 위하여 제거하는 빈발항목모임들속에는 부의 련관규칙을 발굴하는데 유용하게 쓰이는 항목모임들이 있게 된다. 그러므로 전통적인 알고리듬으로는 부의 련관규칙발굴에 필요한 흥미있는 부의 항목모임들을 찾아낼수 없다. 론문에서 제기하는 알고리듬은 흥미있는 정의 항목모임들뿐아니라 흥미있는 부의 항목모임들도 모두 탐색하게 함으로써 정 및 부의 련관규칙들을 모두 발굴할수 있는 전제를 마련한다.

우리는 이미 흥미있는 정의 항목모임과 부의 항목모임에 대한 명백한 정의를 제시하고 그에 기초하여 전통적인 련관규칙발굴알고리듬인 Apriori가 생성된 후보항목모임들(C_k)중에서 앞선 단계의 빈발항목모임족(L_{k-1})에 속하지 않는 빈발항목모임들은 제거하기 때문에 부의 련관규칙발굴에 적합하지 않다는것을 론의하였다.[1, 2]

또한 정 및 부의 항목모임들의 수가 대단히 크기때문에 탐색공간문제가 제기되며 이로부터 련관규칙발굴에 필요한 흥미있는 항목모임들만을 식별하는것이 중요하다는것을 밝혔다.[1, 2]

부의 련관규칙의 한가지 종류인 드문련관규칙발굴과 관련한 선행연구들이 제기되였지만 특정의 경우만을 고찰한데로부터 흥미있는 항목모임들의 정의와 탐색문제를 론의하지 못하였다.[3, 4]

우리가 앞서 제시한 정의 련관규칙은 $X \rightarrow Y$형태이며 부의 련관규칙은 $X \rightarrow \neg Y$형태이다. 그리고 흥미없는 항목모임들은 자료기지에서 흥미있는 정 및 부의 항목모임들을 둘다 배제하는 임의의 항목모임으로 된다. 이러한 항목모임들은 발굴시 탐색되는 공간을 줄일수 있게 잘라버릴 필요가 있다. 다시말하여 흥미없는 련관규칙과 관련되는 빈발항목모임들이 많이 존재한다. 항목모임들중에서 흥미있는 정 및 부의 항목모임들만을 추출한다면 탐색공간은 극히 감소될수 있다.

1. 알고리듬의 설계

알고리듬에서 주어진 자료는 자료기지 *D*, 최소지지도 *minsprt*, 최소믿음도 *minconf*, 최소흥미도 *mininterest*이다. 알고리듬에서 결과로 얻어지는 자료는 흥미있는 정의 항목모임족 *PS*, 흥미있는 부의 항목모임족 *NS*이다.

알고리듬의 실행과정에 매 단계($k>0$)에서 빈발항목모임족 $Freq_k$가 생성되여야 하며 그에 기초하여 정의 항목모임족 P_k와 부의 항목모임족 N_k가 생성되여야 한다. 여기서 P_k는 전통적인 련관규칙발굴의 알고리듬에서 사용하던 빈발항목모임족 L_k와 동일하다.

이제 $Temp_k=P_k\cup N_k$라고 하자. 그러면 전통적인 련관규칙발굴의 알고리듬에서 사용하던 후보모임 C_k는 $Temp_k$의 부분모임으로서 그안의 매 원소는 P_{k-1}의 원소로 되는 적어도 하나의 부분모임을 포함해야 한다. 결국 $Temp_k$는 자료기지안의 모든 $k-$항목모임들의 족으로서 $Temp_k$안의 매 항목모임은 $Freq_i(1\le i\le k-1)$안의 어떤 2개의 빈발항목모임들의 합이다. 즉 $Freq_{i_0}$ 안의 항목모임 A와 $Freq_{i_1}$ 안의 항목모임 $B(1\le i_0,\ i_1\le k-1)$에 대하여 $A\cup B$가 $k-$항목모임이면 $A\cup B$는 Tem_k안에 추가된다. 그리고 $Temp_k$안의 매 항목모임에 대하여 자료기지 D안에서 지지도를 계수하여야 한다.

다음으로 만일 P_k의 항목모임 $i=X\cup Y$가 임의의 X와 Y에 대하여 식 $|sprt(X\cup Y)-sprt(X)sprt(Y)|<mininterest$를 만족시킨다면 i는 흥미없는 빈발항목모임이며 그것은 P_k에서 잘라버려야 한다. P_k에서 모든 흥미없는 항목모임들을 제거한 다음 모임 P_k를 PS에 추가하면 된다. 마찬가지로 만일 N_k의 항목모임 $i=X\cup Y$가 임의의 X와 Y에 대하여 식 $|sprt(X\cup Y)-sprt(X)sprt(Y)|<mininterest$를 만족시킨다면 i는 흥미없는 빈발항목모임이며 그것은 N_k에서 제거하여야 한다. N_k에서 모든 흥미없는 항목모임들을 제거한 다음 모임 N_k를 NS에 추가하면 된다.

우와 같은 단계의 순환의 끝조건은 $P_k\ne\varnothing$, $N_k\ne\varnothing$이다. 흥미있는 정의 항목모임들과 부의 항목모임들을 각각 PS와 NS로 출력한다.

알고리듬은 다음과 같다.

흥미있는 항목모임의 탐색알고리듬(Searching of Interesting Itemsets)

입력자료: D(자료기지),

$minsprt$(최소지지도), $minconf$(최소믿음도), $mininterest$(최소흥미척도)

출력자료: PS(흥미있는 정의 항목모임족), NS(흥미있는 부의 항목모임족)

(1) $PS=\varnothing$; $NS=\varnothing$;

(2) $Freq_1$=1-빈발항목모임들의 족; // D의 첫 통과

k=1;

// 흥미있는 모든 정 및 부의 $k-$항목모임들을 생성

(3) do {

k++;

//① 가능한 $k-$항목모임들을 생성

$Temp_k=\{A\cup B|\ A\in Freq_{i_0},\ B\in Freq_{i_1},\ (1\le i_0,\ i_1\le k-1),\ |A\cup B|=k\}$;

//② D의 트랜잭션 t에 포함되여있는 $k-$항목모임들을 계수

for $\forall t\in D$ *do* {

$Temt=\{k\text{-}itemset|\ k\text{-}itemset\subseteq t,\ k\text{-}itemset\in Temp_k\}$;

for $\forall itemset\in Temt$ do

$itemset.count=itemset.count+1$;

}

//③ k－후보항목모임과 k－빈발항목모임에 기초하여 k－정 및 부의 항목모임들을 생성

C_k={*k-itemset*| *k-itemset*∈$Temp_k$, ∃*Sub*⊂*k-itemset*: *Sub*∈P_{k-1}}; // k－후보항목모임

$Freq_k$={*c*| *c*∈C_k, *sprt*(*c*)=*c.count*/|*D*| >=*minsprt*}; // k－빈발항목모임

P_k= $Freq_k$, // k－정의 항목모임

NN_k= $Temp_k$ －P_k

//④ 흥미없는 k－정의 항목모임자르기

for ∀ *itemset* ∈P_k do

if *itemset*: 흥미없는 항목모임 then P_k=P_k – {*itemset*};

PS=*PS*∪ P_k ;

//⑤ 흥미없는 k－부의 항목모임자르기

N_k= {*itemset* ∈NN_k, *itemset*: 부의 항목모임}; // k－부의 항목모임

for ∀ *itemset* ∈N_k do

if *itemset*: 흥미없는 항목모임 then N_k=N_k – {*itemset*};

NS=*NS*∪N_k;

} while ($P_{k-1} \neq \varnothing$, $N_{k-1} \neq \varnothing$);

// 결과출력

(4) output *PS*, *NS*;

우의 알고리듬을 리용하여 지수적크기를 가진 항목모임들의 탐색공간에서 흥미없는 항목모임들을 제거하면 탐색공간은 현저히 감소된다.

2. 알고리듬의 적용

알고리듬의 적용과정을 실례를 통하여 보자. 표에 10개의 트랜잭션으로 이루어진 자료기지가 주어졌다. 5개의 항목들인 A, B, C, D, E, F가 있다. 그리고 *minsprt*=0.3, *mininterest*=0.07로 가정하자.

자료기지의 첫 통과로 매 항목들에 대한 지지도를 구하면 A: 5/10=0.5, B: 7/10=0.7, C: 6/10=0.6, D: 6/10=0.6, E: 3/10=0.3, F: 5/10=0.5이다.

이로부터 *minsprt*=0.3이므로 모든 항목들이 1－빈발항목모임이다. 따라서 $Freq_1$={A, B, C, D, E, F}이다.

$Freq_1$에 의해 생성되는 $Temp_2$={AB, AC, AD, AE, AF, BC, BD, BE, BF, CD, CE, CF, DE, DF, EF}이며 *minsprt*=0.3에 기초하여 *Temt*, C_2들을 거쳐 생성되는 $Freq_2$=P_2={AB, AC, AD, BC, BD, BF, CD, CF}이다. 그러므로 NN_2= $Temp_2$-P_2={AE, AF, BE, CE, DE, DF, EF}이다.

mininterest=0.07에 기초하여 P_2에서 흥미없는 항목모임들을 제거하기 위하여 매 항목모임의 흥미척도를

표. 실례자료기지

트랜잭션 ID	항목들
T1	A, B, D
T2	A, B, C, D
T3	B, D
T4	B, C, D, E
T5	A, C, E
T6	B, D, F
T7	A, E, F
T8	C, F
T9	B, C, F
T10	A, B, C, D, F

계산하면 BD만이 흥미있고 나머지모임들은 모두 흥미가 없다. 따라서 흥미없는 항목모임들을 *PS*에 추가하기 전에 제거하면 P_2={BD}이다.

마찬가지로 NN_2의 매 항목모임의 흥미척도를 계산하면 BE, DF만이 흥미있고 나머지 모임들은 모두 흥미가 없다. 따라서 흥미없는 항목모임들을 *NS*에 추가하기 전에 제거하면 N_2={BE, DF}이다.

다음순환에서 $Freq_1$과 $Freq_2$에 의해 생성되는 $Temp_3$={ABC, ABD, ABE, ABF, ACD, ACE, ACF, ADE, ADF, AEF, BCD, BCE, BCF, BDE, BDF, BEF, CDE, CDF, CEF, DEF}이며 *minsprt*=0.3에 기초하여 *Temt,* C_3들을 거쳐 생성되는 $Freq_3=P_3$={ABD, BCD}이다.

그러므로 $NN_3=Temp_3-P_3$={ABC, ABE, ABF, ACD, ACE, ACF, ADE, ADF, AEF, BCE, BCF, BDE, BDF, BEF, CDE, CDF, CEF, DEF}이다.

mininterest=0.07에 기초하여 P_3에서 흥미없는 항목모임들을 제거하기 위하여 매 항목모임의 흥미척도를 계산하면 ABD, BCD는 정의로부터 둘 다 흥미있는 정의 항목모임이다. 따라서 P_3={ABD, BCD}이다.

마찬가지로 NN_3의 매 항목모임의 흥미척도를 계산하고 흥미없는 부의 항목모임들을 *NS*에 추가하기 전에 제거하면 N_3={ABE, ADE, BDE, BDF, BEF, CDF, CEF}이다.

다음순환에서 $Freq_1$, $Freq_2$, $Freq_3$에 의해 생성되는 $Temp_4$={ABCD, ABCF, ABDE, ABDF, BCDE, BCDF}이며 *minsprt*=0.3에 기초하여 *Temt*, C_4들을 거쳐 생성되는 $Freq_4=P_4=\varnothing$이다. 그러므로 $NN_4= Temp_4-P_4$={ABCD, ABCF, ABDE, ABDF, BCDE, BCDF}이다.

mininterest=0.07에 기초하여 NN_4에서 흥미없는 항목모임들을 제거하기 위하여 매 항목모임의 흥미척도를 계산하고 흥미없는 항목모임들을 제거하면 N_4={ABCD, ABDE}이다.

최종적으로 생성되는 $Temp_5$={ABCDF}이며 *minsprt*=0.3에 기초하여 $Freq_5=P_5=\varnothing$이다. 그러므로 $NN_5=Temp_5-P_5$={ABCDF}이다. ABCDF는 흥미척도의 계산으로부터 흥미가 없다는 것을 알수 있다. 따라서 NN_5로부터 흥미없는 항목모임을 제거하면 $N_5=\varnothing$이다. 순환은 여기서 끝나며 알고리듬은 다음의 결과를 출력한다.

PS={P2, P3, P4, P5}={BD, ABD, BCD}

NS={N2, N3, N4, N5}={BE, DF, ABE, ADE, BDE, BDF, BEF, CDF, CEF, ABCD, ABDE}

알고리듬의 적용결과로 11개의 빈발항목모임들(1－항목모임들은 제외)중에서 4개만이 흥미있는 정의 항목모임이며 31개의 부의 항목모임들중에서 11개만이 흥미있는 부의 항목모임들로 얻어졌다.

이 알고리듬에서 흥미있는 정 및 부의 항목모임들을 식별할 때 규칙의 믿음도가 *minconf*보다 크거나 같아야 한다는것을 고려하지 않았다. 즉 흥미있는 정 및 부의 항목모임들의 매개에 대하여 최소믿음도조건을 고려하여야 한다. 이 조건을 고려하면 탐색공간은 더 줄어들게 된다.

알고리듬의 적용과정을 통하여 흥미있는 빈발항목모임들이 탐색될 때 어떤 하나의 흥미있는 빈발항목모임이 직접 제거된다면 그것은 빈발항목모임들속에서 더는 발생할수 없지만 반대로 그것은 흥미있는 비빈발항목모임들속에 발생할수 있다는것을 알수 있다. 그다음 만일 그 항목모임이 흥미있는 비빈발항목모임들이 탐색될 때 잘리운다면 그것은 탐색되는 공간의 나머지에서 제거된다. 한편 흥미있는 비빈발항목모임들이 탐색될 때 어

떤 하나의 비빈발항목모임이 제거된다면 그것은 흥미있는 빈발항목모임들에 대한 탐색에 영향을 주지 않는다는것도 알수 있다.

3. 결 론

정 및 부의 련관규칙들을 다 발굴하자면 빈발항목모임들뿐아니라 비빈발항목모임들도 탐색하여야 한다. 빈발항목모임들을 식별하는것은 그자체가 주어진 자료기지안에 있는 모든 가능한 항목과 항목모임들로 이루어지는 지수적공간을 탐색하는것으로 되며 한편 가능한 비빈발항목모임들의 수는 빈발항목모임들의 수보다 더 크다. 즉 가능한 정 및 부의 항목모임들의 량은 거의 두배로 된다.

론문에서는 정 및 부의 련관규칙발굴의 첫 단계에서 해결해야 할 흥미있는 정 및 부의 항목모임들을 탐색하는 알고리듬을 제기함으로써 련관규칙발굴의 탐색공간을 줄이는 문제를 취급하였다. 그리고 실례를 통하여 흥미있는 항목모임들로 련관규칙발굴의 탐색공간이 현저히 줄어든다는것을 보여주었다.

참 고 문 헌

[1] K. Hyeok et al.; IJTPC, 12, 12, 13, 2016.

[2] K. Hyeok et al.; IJTPC, 10, 12, 1, 2015.

[3] M. Almasi et al.; ELSEVIER: Knowledge-Based Systems, 89, 366, 2015.

[4] S. Jain et al.; Proceedings of Third IRF International Conference, ISBN: 978-93, 34, 2015.

주체107(2018)년 6월 5일 원고접수

An Algorithm for Searching Itemsets of Interest

Pae Chol Jin, Kong Hye Ok

This paper presents an algorithm for searching positive and negative itemsets of interest, which is the first step for the positive and negative association rules mining, to reduce the searched space of it. And, we show the searched space of itemsets of interest is much reduced through an example.

Key words: positive and negative itemsets, negative association rule