

# 비연계 DB 테이블상에서의 데이터 추출을 위한 규칙 기반의 데이터 마이닝 기법

## A Rule-Based Data Mining Method among the Unrelated DataBase Table

김 찬 일, 조 대 호  
성균관대학교 전기전자 및 컴퓨터공학부  
(E-mail : ili72@peter.skku.ac.kr)

### ABSTRACT

데이터 마이닝란 대량의 실제 데이터에서 묵시적이고 잠재적으로 유용한 정보를 추출하는 작업이다. 본 논문에서 서로 관계가 정의되지 않은 데이터베이스의 각 테이블간에서 필요한 정보를 추출 또는 가공하기 위해 데이터 마이닝 기법을 사용한다. 마이닝 기법인 연관 규칙은 어떤 사건이 일어나면 다른 사건이 일어나는 관련성을 의미하는 것이고, 제시된 규칙 기반의 데이터 마이닝 기법은 연관 규칙의 한 분야로서 데이터를 규칙 맞게 분류하는 기법이다. 이런 마이닝 기법을 구현하기 위해 인공 지능 분야의 규칙 기반의 전문가 시스템을 사용하였고, 실 시스템인 GDS(Grating automatic Drawing System)에 적용하였다.

### 1. 서론

데이터 마이닝은 대량의 실제 데이터로부터 이전에 잘 알려지지 않는 묵시적이고 잠재적으로 유용한 정보를 추출하는 작업이라 정의한다.[1] 이런 데이터 마이닝 기법을 이용하여 테이블간의 관계를 정의하고 데이터베이스의 필요한 값을 찾는 방법을 제시한다.

데이터베이스에서 값을 찾기 위해서는 찾고자하는 값에 맞는 조건인 Query를 주며 각각의 테이블에서 Query에 맞는 값들을 찾아온다. 여기서 값을 찾을 때, 각 테이블의 에트리뷰트는 서로의 테이블과 어떤 관계를 가지고 있다. 이런 관계는 초창기 데이터베이스를 설계할

때 정의되어야 하고 그 정의된 설계에 따라 데이터베이스는 구성되어진다. 그러나, 수정 또는 유지 보수 시 추가되는 테이블들이 종종 있을 수 있고 테이블의 에트리뷰트들이 바뀌거나 재구성이 되는 경우가 있다. 무엇보다 주목해야 하는 점은 관계가 있지 않은 다른 테이블을 통해서 필요한 값을 찾는 경우이다. 이런 경우에는 데이터베이스 설계를 바꾸던지 아니면 새로 생긴 테이블의 에트리뷰트를 기존에 있는 테이블에 첨가하여 데이터베이스를 재 설계해야 한다.

본 연구에서는 데이터베이스 관계를 시스템에서 제어 할 수 있도록 관계가 정의되지 않은 테이블 상에서도 필요한 데이터를 찾을 수 있는 방법을 제시한다.

## 2. 연관 규칙과 전문가 시스템

### 2.1 연관 규칙

데이터 마이닝은 결정 트리 뉴럴 네트워크, 유전자 알고리즘 그리고 연관 등이 연구되고 있다.[2] 이 중 연관 규칙이란 어떤 집합이 서로 소인 다른 집합들과 일어나는 관련성을 의미한다. 즉 주어진 전체 S를 집합  $S = \{S-1, \dots, S-n\}$ 라 하였을 때 서로 소의 관계인 두 개의 부분집합  $A = \{S-11, S-12, \dots, S-1m\}$ 과  $B = \{S-21, S-22, \dots, S-2k\}$ 라 할 때 A에서 B로 일어나는 관련성을 "A->B"로 표시된다. 이때 집합 A는 집합 B를 연관이라고 정의한다. 즉 A는 B를 유도한다고 할수 있다.

### 2.2 전문가 시스템

전문가 시스템은 적용 영역의 전문가들이 가지고 있는 전문 지식을 지식베이스로 구축하여 저장함으로써 컴퓨터가 전문가의 기능을 인간과 같은 논리적인 사고로 대신 수행케 하는 시스템이다.

## 3. 연계와 비연계 DB 테이블 구성

### 3.1 연계 DB 테이블에서 데이터 추출

연계 DB 테이블에서 데이터를 추출하기 위해서는 각 테이블의 에트리뷰트를 기반으로 테이블들간의 관계를 맺고 이 조건으로 각 테이블에서 필요한 데이터를 추출하였다.

즉 다음과 같은 3개의 테이블이 있다.

Item Table						DrawCode Table				
IT	DC	BM	LEG	WID	EC	DC	CC	TC	OD	JN
I-1	SKU	BM-3	5	4	KCI	SKU	CC-2	TC-2	KIM	Test
I-2	SKU	BM-1	3	2	LMR	SUN	CC-1	TC-1	LEE	Test

표 1

표 2

Item당 너비 비례의 Clip의 중량을 구하기 위

ClipCode Table					추출된 데이터				
CC	CT				IT	LEG	WID	CT	
CC-1	10				I-1	5	4	15	
CC-2	15				I-2	3	2	15	

해서는 다음과 같은 Query를 주어 필요한 값들을 찾아온다.

표 3

SELECT : LEG, WID, CT

표 4

```
FROM : Item, DrawCode, ClipCode
WHERE : Item.DC=DrawCode.DC AND
DrawCode.CC=ClipCode.CC
```

우리는 3개의 테이블을 조인하여 데이터를 추출한 결과는 표 4이다.

이렇게 찾은 데이터를 연산하면

I-1의 Clip무게 :  $5 \times 4 \times 15 = 300$

I-2의 Clip무게 :  $3 \times 2 \times 15 = 90$

라는 값을 얻을 수 있다.

여기서 Item 테이블은 DrawCode 테이블과 DC라는 에트리뷰트로 연결되어 있고 DrawCode 테이블은 ClipCode 테이블과 CC라는 에트리뷰트로 연결되어 있다. 즉 Item 테이블과 ClipCode 테이블 사이에 DrawCode 테이블이 연결되어 있는 것이다.

연계 DB 테이블에서는 데이터베이스 설계자가 테이블의 관계 관련성을 알고 초창기 데이터베이스 설계시 정적으로 연관 관계가 미리 정의된 데이터베이스이기 때문에 테이블 추가 또는 수정에 어려운 점이 있다. 이런 관련성을 동적으로 연결할 수 있어야 수정의 용의성과 유지보수의 편리성을 데이터베이스에 제공할 수 있다.

본 연구에서 DrawCode 테이블 같은 중간 단계 테이블(1개 이상의 중간 테이블) 없이 테이블과 테이블 사이에서 필요한 정보를 찾는 방법을 제시 하고자 한다.

### 3.2 비연계 DB 테이블에서 데이터 추출

중간 단계 테이블이 사라지면 이 테이블이 하는 역할 즉 테이블과 테이블을 연결 시켜주는 관계를 정의하는 역할을 데이터 마이닝 기법으로 정의하도록 제시하고 이 기능을 수행할 시스템으로 전문가 시스템을 사용한다.

기존의 3개의 데이터베이스 테이블을 가진 시스템은 다음과 같습니다.

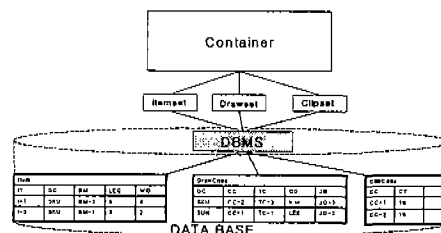


그림 2

새롭게 제시된 시스템의 형태는 다음과 같습

니다.

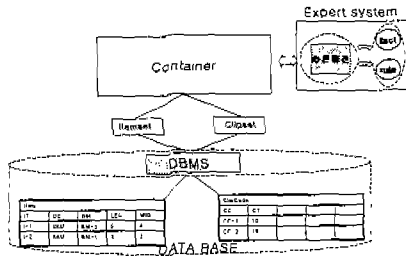


그림 3

Item 테이블에서 DC 값을 추출한 다음 전문가 시스템의 추론으로 ClipCode와 한 애틀리뷰트를 연관이 있는 것을 추론하고 ClipCode 테이블의 CT 값을 찾아 LEG와 WID 값을 곱하여 무게를 계산한다.

적용되는 규칙을 다음과 같이 정의 됩니다.

Mapping Table : DC=SKU -> CC=CC-2  
DC=SUN -> CC=CC-1

Fact : IT=I-1, CT

Rule 1

If : IT에 값이 있다.

Add : LEG=5, WID=4, DC=SKU

Fact : IT=I-1, CT, LEG=5, WID=4, DC=SKU

Rule 2

If : CT 존재하고 DC값이 있다.

Mapping : CC=CC-2

Delete : DC 제거

Fact : IT=I-1, CT, LEG=5, WID=4, CC=CC-2

Rule 3

If : CT 존재 CC, LEG, WID 값이 있다.

Add : CT=15

그러나, 여기서 우리는 한가지 유의 할 점이 있다. Item 테이블은 ClipCode 테이블 데이터만을 추출하지 않는다는 점이다. 즉 다른 테이블들에 있는 데이터를 추출하기를 원할 수도 있다.

그래서 다음과 같이 3개의 테이블들을 첨가한다.

Item Table					
IT	DC	BM	LEG	WID	EC
I-1	SKU	BM-3	5	4	KCI
I-2	SKU	BM-1	3	2	LMR

표 5

DrawCode Table				
DC	CC	TC	OD	JN
SKU	CC-2	TC-2	KIM	Test
SUN	CC-1	TC-1	LEE	Test

표 6

ClipCode Table	
CC	CT
CC-1	10
CC-2	15

표 7

BMList Table	
BM	GC
BM-1	GC-1
BM-2	GC-1

표 8

GratCode Table				
GC	GT			
GC-1	10			
GC-2	15			

표 9

ToeCode Table				
TC	TT			
TC-1	2.5			
TC-2	3.5			

표 10

여기서 우리는 Item 테이블에 관련하여 3개의 테이블에서 데이터를 추출 할 수 있다는 것을 알 수 있고 다음과 같은 결과를 얻을 수 있다.

Item당 너비에 대한 Clip의 무게를 계산  
SELECT : CT  
FROM : Item, DrawCode, ClipCode  
WHERE : Item.DC=DrawCode.DC AND  
DrawCode.CC=ClipCode.CC

Item당 너비에 대한 무게 계산

SELECT : GT  
FROM : Item, BMList, GratCode  
WHERE : Item.BM=BMList.BM AND  
BMList.GT=GratCode.GT

ClipCode Table				
IT	CT			
I-1	15			
I-2	15			

표 11

GratCode Table				
IT	GT			
IT-1	10			
IT-2	15			

표 12

SELECT : TT  
FROM : Item, DrawCode, ToeCode  
WHERE : Item.DC=DrawCode.DC AND  
DrawCode.TT=ToeCode.TT

Item당 Toe 무게 계산

ToeCode Table				
IT	TT			
I-1	3.5			
I-2	3.5			

표 13

Item 테이블은 주어진 상황에 따라 각각의 다른 테이블에서 데이터를 추출할 수가 있는 것이다. 즉 비연계 DB 테이블에서 데이터를 추출하기 위해서 묵시적이고 잠재적으로 유용한 정보를 추출하는 작업이 필요하게 된다.

본 논문에서 규칙 기반의 연관 패턴을 제시하고자 한다.

## 4. 규칙 기반 패턴

### 4.1 일반화된 연관 규칙 형태

규칙 기반 패턴은 데이터 마이닝 기법인 연관 기법 중 한 분야이다. 연관 규칙에 대한 연구는 IBM Almaden 연구소에서 QUEST project로서 활발히 진행되고 있다.[3] 기존의

연관 규칙은 데이터의 속성으로 분류 (taxonomy)한다. 예로서 다음과 같은 형태로 재구성한다.

F-T : CT, GT, TT의 값을 얻기 위해 처음으로 알아야 하는 값의 집합

S-T : F-T의 값에 의해 두 번째로 알아야 하는 값의 집합

T-T : S-T의 값에 의해 세 번째 값을 알아야 하는 값의 집합

F_T	S_T	T_T	Goal		
Item	DC	CC	CT		
Item	DC	GC	GT		

표 14

CT를 찾기 위해서 Item, DrawCode, ClipCode의 테이블들, TT를 찾기 위해서는 Item, DrawCode, ToeCode 의 테이블들과 GT 를 찾기 위해서 Item, BMList, GratCode 의 테이블들이 연관관계를 가진다는 사실을 알 수가 있다.

이런 연관 관계를 다음과 같은 그림으로 표시하였다.

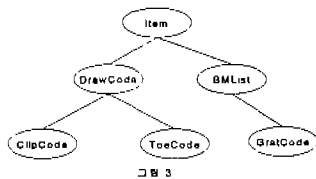


그림 3

## 4.2 규칙 기반 패턴 형태

4.1절에서 설명한 방식으로 연관 규칙을 제시 하면 우리가 제시하고자 하는 비연계 DB에서 데이터를 추출 할 수가 없다. 그래서 규칙 기반 패턴으로 테이블 사이의 연관 관계를 규칙으로 연결하는 방법을 제시하고자 한다.

위의 예를 규칙 기반 패턴으로 연결하면 다음과 같은 형태로 형성이 된다.

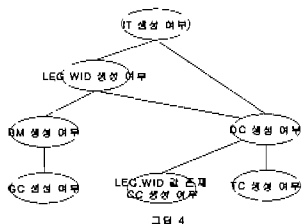


그림 4

먼저 IT의 생성 여부를 인식하고 이 IT의 값으로 LEG, WID의 값이 생성되는지 추론하여 생성된 것을 인지하고 BM생성 여부를 추론한

다. 이 추론된 결과가 LEG, WID 존재 여부와 GC 생성 여부를 추론하여 우리가 마지막으로 찾고자 하는 Item당 너비에 대한 무게 계산을 얻을 수 있는 것이다. Item당 너비에 대한 Clip의 무게를 계산, Item당 Toe 무게 계산 값도 위와 같은 방법으로 추론하면 그에 맞는 값을 찾을 수 있다.

대용량의 데이터 사이의 묵시적이고 잠재적으로 유용한 정보를 규칙으로 구성하면 이 묵시적이고 잠재적으로 유용한 정보를 사용해 찾아지는 데이터 추출은 비연계 DB 테이블에서도 데이터를 추출 할 수 있고 비연계 DB 테이블로 구성되기 때문에 수정의 용의성과 유지보수의 편리성도 제공이 되는 것이다.

## 5. 결론

본 연구는 첫째 비연계 테이블들 사이에서도 묵시적이고 잠재적으로 유용한 정보를 찾는 방법인 규칙 기반 패턴을 제시함으로써 테이블 사이의 관계(relation)에 상관없이 필요한 데이터를 추출하는 것과 비연계 DB 테이블로 구성되어진 데이터베이스에서 데이터를 추출하기 때문에 수정의 용의성과 유지보수의 편리성 구축의 간편함을 제공한다. 또 이런 기능을 수행할 목적으로 전문가 시스템을 사용하기 때문에 데이터베이스에 있는 데이터 제어 부분을 시스템에서 다루어 데이터 제어를 쉽게 한다.

### 참고 문헌

- [1] 김정자, 이도현 "데이터 마이닝 기술 및 연구 동향" 정보과학회지 16권 제9호 1998.9
- [2] Groth Robert Data Mining Building Competitive Advantage Prentice-Hall Inc. 1999
- [3] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules in large databases", In *Proceedings of ACM SIGMOD Conference on Management of Data*, Washington D.C., 1993.5
- [4] M. Houtsma and A. Swami, "Set-Oriented Mining for association rules", IBM Research Report, RJ 9567 (83573) 1993.8
- [5] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In *Proceeding of the 20th VLDB Conference*, Santiago, Chile, 1994.9
- [6] P. H. Winston, "Artificial Intelligence : The Third Edition", Addison Wesley, 1992, chap 7
- [7] Stonebraker M., Woodfill, J., and Andersen, E. , "Implementation of Rule in Relational Data Base Systems", Database Engineering Vol. 6 No.4, 1983, pp.65-74

[8] R. Srikant and R. Agrawal, "Mining Generalized Association Rules" In *Proceeding of the 21th VLDB Conference*, Zurich,Switzerland, 1995

[9] R. Agrawal and R. Srikant, "Mining sequential patterns", In *Proceedings of the 11th International Conference on Data Engineering*, Taipei, Taiwan, 1996.3

[10] M.-S. Chen, J. Han, and P.S Yu,"Data Mining: An Overview from a Database Perspective", *IEEE Transaction on Knowledge and Data Engineering*, Vol. 8, No. 6 1996.10