Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 2, pp. 279 ~ 287

Association Rule Mining by Environmental Data Fusion

Kwang-Hyun Cho[1], Hee-Chang Park[2]

Abstract

Data fusion is the process of combining multiple data in order to produce information of tactical value to the user. Data fusion is generally defined as the use of techniques that combine data from multiple sources and gather that information in order to achieve inferences. Data fusion is also called data combination or data matching. Data fusion is divided in five branch types which are exact matching, judgemental matching, probability matching, statistical matching, and data linking.

In this paper, we develop sas macro program for statistical matching which is one of five branch types for data fusion. And then we apply data fusion and association rule techniques to environmental data.

Keywords : Asociation Rule, Data Fusioin, SAS Macro, Statistical Matching

1. 서론

사회지표는 주민들이 생각하는 사회상태를 총체적이고도 집약적으로 나타내는 것으로, 변화하는 역사적 흐름 속에서 우리가 처해 있는 사회적 상태를 종합적이고 집약적으로 나타냄으로써 사회구성원들의 삶의 질을 전반적으로 파악하고 사회변화를 포착할 수 있는 척도이다. 현재 경상남도는 도민들을 대상으로 3년 주기로 매년 설문문항을 다르게 하여 사회지표 조사를 실시하고 있어 도민들의 환경의식에 대한 분석 시 연도별로 각각 분석을 실시해야 함으로써 유기적인 분석이 가능하지 못한 실정이다. 또한, 특정 연도의 사회지표 조사 자료에서는 환경의식 분석에 사용할 환경 관련 문항들이 기타 연도에 비하여 작아 다양한 분석을 실시하지 못하고 있다. 이에 각 연

1) Graduate Student, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : cho1023@changwon.ac.kr

2) Corresponding author : Professor, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : hcpark@changwon.ac.kr

도의 사회지표 조사 자료를 결합하여 하나의 데이터 파일을 만들면 고부가가치의 정보를 획득할 수 있을 것이며, 이를 위하여 사용되어 지는 방법 중의 하나가 데이터 퓨전(data fusion) 방법이다. 데이터 퓨전은 같은 모집단에서 나온 서로 다른 표본들을 포함하는 데이터 셋을 합치는 기법 또는 처리과정으로 정의된다.

본 논문에서는 도민들의 환경 의식을 더욱더 총체적이고 효율적으로 분석하기 위하여 경상남도에서 조사된 사회지표 조사 자료의 환경자료에 대하여 데이터 퓨전에 의한 연관성규칙을 적용하고자 한다. 데이터 퓨전에 대한 국내 연구로는 최기주와 정연식(1998)은 교통정보의 생성을 위하여 링크 통행시간 추정을 하는데 데이터 퓨전을 이용하였고, 손소영과 이성호(2000)는 도로교통사고 자료처리에서의 교통사고 심각도 분류분석에 데이터 퓨전을 적용하였으며, 신형원과 손소영(2000)은 다구찌 디자인을 이용한 데이터 퓨전의 성능 비교에 대하여 연구한 바 있다. 또한 박성원 등(2001)은 데이터 퓨전을 이용하여 얼굴영상 인식 및 인증에 관하여 연구한바 있으며 김호종 등(2003)은 신경망 기반 추천 모델의 성능향상을 위하여 데이터 퓨전을 적용하였다. 각 연도의 환경자료는 데이터 퓨전 기법에 의하여 하나의 데이터 파일로 결합되고 결합된 파일에 대하여 연관성 규칙을 적용한다.

한편, 본 논문에서 적용한 연관성규칙(association rule)은 하나의 거래나 사건에 포함되어 있는 둘이상의 품목들의 경향을 파악해서 상호 관련성을 발견하는 것으로, 둘 또는 그 이상의 품목들 사이의 지지도(support), 신뢰도(confidence), 향상도(lift)를 바탕으로 관련성 여부를 측정한다. 연관 규칙은 탐색적이며, 비목적성 분석이며, 기존의 데이터를 특별한 변형 없이 계산이 용이하게 사용 가능하다는 장점을 가지고 있어 현업에서 많이 활용되고 있다. 연관성 규칙은 Agrawal 등(1993)에 의해 처음 소개된 이후, Agrawal과 Srikant(1994), Park 등(1995), Toivonen(1996), Cheung 등(1996), Sergey 등(1997), Saygin 등(2002) 등 국내외의 많은 학자들에 의해 연구되어지고 있다.

본 논문에서는 환경 자료 중 공통으로 가지는 변수에 개인 식별 가능한 변수가 없기 때문에 통계적 결합(statistical matching) 방법을 사용하여 퓨전을 실시한다. 현재 데이터 퓨전을 위한 소프트웨어는 개발되어 있지 않아 SAS 매크로를 이용하여 통계적 결합 알고리즘을 구현하고, 이를 사회지표 조사의 환경자료에 적용하여 각각의 데이터를 하나의 데이터 파일로 생성한 후 연관성 규칙(association rule)을 적용한다. 본 논문의 2절에서는 데이터 퓨전과 연관성 규칙에 대하여 기술하고, 3절에서는 연구방법에 기술하며, 4절에서는 적용 결과를 제시한 후 5절에서 결론을 맺는다.

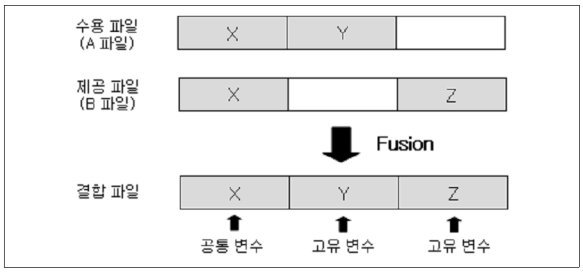
2. 데이터 퓨전과 연관성 규칙

2.1 데이터 퓨전

데이터 퓨전은 그 자체가 하나의 분석이며, 최종 결과라기보다는 통계분석 결과의 질을 높이기 위한 방법이라고 할 수 있다. 즉, 데이터 퓨전을 통해서 얻은 최종 결과에 대한 추가된 정보를 이용함으로써 통계 분석의 질을 향상시킬 수 있다.

서로 다른 두 개의 파일 A와 B를 가정하자. 파일 A와 B에는 X라는 변수가 공통적으로 존재하고 Y변수와 Z변수는 A 파일과 B 파일에 각각 존재한다고 하자. 즉, 파일 A와 B에 공통적으로 X라는 변수가 있고 파일 A에는 Y라는 변수만 존재하며, 파일

B에는 Z라는 변수만 존재한다고 하자. 변수 X, Y, Z로 구성된 파일을 만들기 위하여 <그림 1>과 같이 파일 A와 B를 결합하여 하나의 파일로 만들면 된다. 여기서, 파일 A와 파일 B에 공통적으로 존재하는 변수 X를 공통 변수라고 하고 파일 A 또는 파일 B에서만 존재하는 변수 Y와 변수 Z를 고유 변수라고 한다. 데이터 퓨전의 결과로 생성된 결합 파일은 두 파일의 공통변수 X를 이용하여 파일 B에 존재하는 변수 Z를 파일 A에 추가한 형식으로 나타난다. 여기서, 변수 Z를 수용하는 파일 A를 수용 파일이라 하고, 변수 Z를 제공하는 파일 B를 제공 파일이라고 한다. 파일 A와 B에 의하여 데이터 퓨전을 수행한 후 생성된 파일을 결합 파일이라고 한다.



<그림 1> 데이터 퓨전

영국 National Statistics(2003)에 따르면 데이터 결합(data matching)의 종류는 정확결합(exact matching), 판단 결합(judgemental matching), 확률적 결합(probability matching), 통계적 결합, 데이터 연결(data linking) 등 5가지로 구분된다. 이 논문에서 사용한 통계적 결합은 통계적 결합은 공통으로 가지는 변수는 존재하나 고유 식별 변수처럼 개인 식별 가능한 변수가 없을 때 회귀분석, 로지스틱 회귀분석 등을 사용하여 통계적 방법을 사용하여 데이터 결합하는 방법이다.

2.2 연관성 규칙

Agrawal 등(1993)에 의해 처음 소개된 연관성 규칙은 탐색적이며, 비목적성 분석이며, 기존의 데이터를 특별한 변형 없이 계산이 용이하게 사용 가능하다는 장점을 가지고 있으며, 계산 과정이 길고, 반복된 계산이 많으며, 적절한 품목의 결정이 어렵고, 각 품목의 단위에 따른 표준화가 어렵다는 단점을 아울러 가지고 있다. 연관성 규칙은 이러한 단점에도 불구하고 두 품목간의 관계를 명확히 수치화함으로써 두 개 이상의 품목간의 관련성을 표시하여 주기 때문에 현업에서 많이 활용되고 있다. 연관규칙을 평가하는 기준에는 지지도(support), 신뢰도(confidence), 향상도(lift) 등이 있으며 다음과 같다.

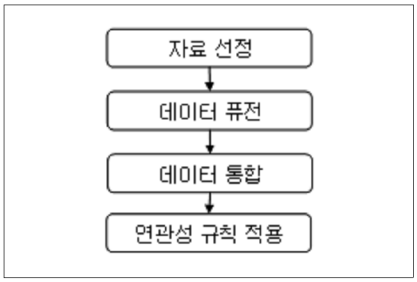
$$\text{지지도 :}\quad S_{(X \Rightarrow Y)} = P(X \cap Y) \tag{2.1}$$

$$\text{신뢰도 :}\quad C_{(X \Rightarrow Y)} = P(Y|X) = \frac{P(X \cap Y)}{P(X)} \tag{2.2}$$

$$\text{향상도 :}\quad L_{(X \Rightarrow Y)} = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)} \tag{2.3}$$

3. 연구 방법

본 논문의 연구 방법은 <그림 2>와 같다.



<그림 2> 연구 방법

3.1 자료 선정

본 논문에서는 2001년 조사된 사회지표 조사와 2002년 조사된 사회지표 조사의 환경자료에 대하여 데이터 퓨전 기법을 적용한다(박희창과 조광현, 2005). 각 조사 자료에 대해서는 공통으로 가지는 변수에 개인 식별 가능한 변수가 없기 때문에 통계적 결합 방법을 사용하여 퓨전을 실시한다. 각각 데이터는 약 10,000건이며 2001년 사회지표 조사에서는 환경관련 6문항과 인구통계학적 속성 5문항으로 구성되어 있고, 2002년 사회지표 조사에서는 2001년 자료 중 인구통계학적 속성 5문항만으로 구성되어 있다. 2001년 조사된 사회지표 조사에 대한 데이터 퓨전 자료는 <표 1>과 같다.

<표 1> 2001년 사회지표 조사 자료 구조

변수 구분	변수명	변수형
환경관련문항	지역의 상수도 환경오염도	연속형
	지역의 하수도 환경오염도	연속형
	지역의 소음진동 환경오염도	연속형
	지역의 악취 환경오염도	연속형
	지역의 대기 환경오염도	연속형
	지역의 토양 환경오염도	연속형
인구통계학적 속성 문항	연령	연속형
	주관적 사회계층	연속형
	학력	범주형
	성별	범주형
	결혼유무	범주형

3.2 데이터 퓨전

본 논문에서 사용한 통계적 결합에 의한 데이터 퓨전 SAS 매크로 프로그램은 다음과 같다.

```
/* data loading */
data data_1; set data.data_2001; run;
data data_2; set data.data_2002; run;
/* macro for continuos matching variable(continuous common variable) */
%macro ContinuosFusion_step1;
/* macro for categorical common variable */
%macro ContinuousFusion_step2;
/* macro for continuous common variable(no use in regression) */
%macro ContinuousFusion_step3;
/* macro for fusion of Continuous variable */
%macro ContinuousFusion_step4;
```

[단계 1] 데이터 준비

수용파일과 제공파일을 지정한다. 수용파일은 data_2002 자료이고 제공파일은 data_2001 이다. 본 논문에서는 결합하고자 하는 변수가 모두 연속형이므로 연속형 문항의 데이터 결합 매크로를 적용하였다.

[단계 2] 연속형 공통변수에 의한 결합 매크로

제공 파일에서의 회귀분석 실시하여 회귀 계수와 추정된 회귀점수를 데이터 셋에 저장한 후 수용 파일에 추정된 회귀식을 적용하여 회귀점수를 계산하여 계산된 회귀점수를 새로운 변수에 저장한다. 회귀점수가 추정된 데이터 셋의 카티션 프러덕트에 대해 공통변수 별로 계산된 회귀점수의 차이를 가지는 레코드들을 계산하고 이로부터 결합변수의 평균을 계산하여 결합한다.

[단계 3] 범주형 공통변수에 의한 결합 매크로

공통변수의 범주 비교한다. 범주 비교 값 계산하여 가장 최소값을 가지는 해당 레코드의 평균값을 계산하여 결합한다.

[단계 4] 회귀분석에서 제외된 연속형 공통변수에 의한 결합 매크로

수용파일에서 표준화 값 계산한다. 제공파일에서의 표준화 값 계산 한다. 표준화된 값의 차이 계산하여 가장 최소값을 가지는 해당 레코드의 평균값을 계산하여 결합한다.

[단계 5] 연속형 결합변수의 최종 결합 매크로

연속형 결합변수의 최종 결합 매크로이다. 연속형, 범주형, 회귀분석에서 제외된 연속형의 3가지의 결합 결과를 통합한다. 각 결합 결과의 평균값을 계산하여 최종적으로 결합한다.

본 논문에서는 6개의 연속형 문항을 결합하므로 [단계 2]에서 [단계 5]의 과정을 6번 반복하여 최종 결합파일을 생성한다.

3.3 데이터 통합

연관성 규칙 적용을 위하여 데이터 퓨전에 의해 생성된 결합파일과 2001년 조사된 사회지표 조사 자료를 통합한다. 데이터 퓨전에 의해 생성된 결합파일의 레코드 수는 9877개이고 2001년 조사된 사회지표 조사 자료의 레코드 수는 9,999개 이므로 최종 통합된 자료의 레코드 수는 19,876개이다.

3.4 연관성 규칙 적용

1) 변수선정 : 연관성 규칙을 생성하기 위하여 전항변수와 후항변수를 선정한다. 후항변수에는 환경관련 문항을 선정하고 전항변수에는 인구통계학적 문항을 선정한다.

2) 전항변수 개수 결정 : 환경관련 문항에 대하여 인구통계학 문항, 구분문항간의 관련성을 알아보기 위하여 전항변수의 개수를 결정해야 한다. 전항변수의 개수가 적으면 연관성 규칙 모형의 정확도는 증가하나 의미 있는 규칙을 찾아내지 못할 수 있고, 전항변수의 개수가 많으면 많은 규칙을 생성하나 연관성 규칙 모형의 정확도가 감소할 수 있다. 연관성 규칙의 신뢰도와 규칙 생성을 고려하여 전항변수의 개수를 2로 설정하였다.

3) 최소지지도, 신뢰도 결정 : 연관성 규칙 생성에 있어 최소지지도와 신뢰도를 결정해야 한다. 최소지지도와 신뢰도를 낮게 결정하면 연관성 규칙의 생성이 많아지나 의미 없는 규칙이 생성될 수 있고, 최소지지도와 신뢰도를 높게 하면 의미 있는 규칙을 찾아내지 못하게 되는 경우도 있다. 이에 최소지지도와 신뢰도를 조절하여 규칙을 생성한 결과 최소지지도 10, 신뢰도를 80으로 결정했을 때 의미 있는 규칙을 발견할 수 있었다.

4. 적용 결과

데이터 퓨전 기법을 이용한 환경자료의 연관성 규칙 적용 결과 지역의 하수도 환경오염도를 제외한 5개 문항에 대하여 의미 있는 규칙을 발견할 수 있었다. 연관성 규칙 적용 시 각 환경오염도 문항에 대하여 부정적인 시각을 나타내는 응답 결과에 대한 규칙만을 기술하였다.

1) 지역의 상수도 환경오염도

지역의 상수도 환경오염도의 연관성 규칙 결과는 <표 2>와 같다. 지역의 상수도 환경오염도에 대한 연관성 규칙 적용 결과 총 10개의 규칙이 생성되었다. 연령이 평균 이하, 결혼 유무가 미혼, 학력이 대제 이상의 응답자들은 지역의 상수도 환경오염도에 대하여 부정적인 시각을 가지고 있는 것으로 나타났다. 연령이 평균 이하이면서 결혼유무가 미혼, 학력이 대제이상, 성별이 여성, 주관적 사회계층이 중하류층 이하인 응답자들은 지역의 상수도 환경오염도에 대하여 부정적인 시각을 가지고 있는 것으로 나타났다. 또한 결혼유무가 미혼이면서 성별이 여성, 주관적 사회계층이 중하류층 이하인 응답자들은 지역의 상수도 환경오염도에 대하여 부정적인 시각을 가지고 있는 것으로 나타났다.

<표 2> 지역의 상수도 환경오염도 연관성 규칙 적용 결과

규칙	지지도	신뢰도	전항값 1	전항값 2
1	47.31	83.42	연령 = 평균이하	
2	17.47	81.01	결혼유무 = 미혼	
3	22.75	88.35	학력 = 대제이상	
4	17.20	81.52	연령 = 평균이하	결혼유무 = 미혼
5	19.90	88.90	연령 = 평균이하	학력 = 대제이상
6	21.54	82.81	연령 = 평균이하	성별 = 여성
7	27.81	85.31	연령 = 평균이하	주관적 사회계층 = 중하류층 이하
8	10.36	80.31	결혼유무 = 미혼	성별 = 여성
9	10.08	84.03	결혼유무 = 미혼	주관적 사회계층 = 중하류층 이하

2) 지역의 소음진동 환경오염도

지역의 소음진동 환경오염도의 연관성 규칙 결과는 <표 3>과 같다.

<표 3> 지역의 소음진동 환경오염도 연관성 규칙 적용 결과

규칙	지지도	신뢰도	전항값 1	전항값 2
1	41.16	80.05	학력 = 고졸이하	
2	42.42	82.50	결혼유무 = 기혼	

지역의 소음진동 환경오염도에 대한 연관성 규칙 적용 결과 총 2개의 규칙이 생성되었다. 학력이 고졸이하, 결혼유무가 기혼인 응답자들은 지역의 소음진동 환경오염도에 대하여 부정적인 시각을 가지고 있는 것으로 나타났다.

3) 지역의 악취 환경오염도

지역의 악취 환경오염도의 연관성 규칙 결과는 <표 4>와 같다.

<표 4> 지역의 악취 환경오염도 연관성 규칙 적용 결과

규칙	지지도	신뢰도	전항값 1	전항값 2
1	10.86	86.38	주관적 사회계층 = 중하류층 이하	학력 = 대제이상
2	10.27	85.58	주관적 사회계층 = 중하류층 이하	결혼유무 = 미혼
3	28.54	87.54	주관적 사회계층 = 중하류층 이하	연령 = 평균이하
4	28.64	83.44	연령 = 평균이하	학력 = 고졸이하

지역의 악취 환경오염도에 대한 연관성 규칙 적용 결과 총 4개의 규칙이 생성되었다. 주관적 사회계층이 중하류층이면서 학력이 대제이상, 결혼유무가 미혼, 연령이 평균이하, 학력이 고졸이하인 응답자들은 지역의 악취 환경오염도에 대하여 부정적인 시각을 가지고 있는 것으로 나타났다.

4) 지역의 대기 환경오염도

지역의 대기 환경오염도의 연관성 규칙 결과는 <표 5>와 같다.

<표 5> 지역의 대기 환경오염도 연관성 규칙 적용 결과

규칙	지지도	신뢰도	전항값 1	전항값 2
1	21.12	82.04	학력 = 대제이상	
2	10.39	82.32	학력 = 대제이상	주관적 사회계층 = 중하류층 이하
3	11.84	82.57	학력 = 대제이상	결혼유무 = 기혼
4	18.36	82.02	학력 = 대제이상	연령 = 평균이하
5	28.63	80.39	연령 = 평균이하	결혼유무 = 기혼

지역의 대기 환경오염도에 대한 연관성 규칙 적용 결과 총 5개의 규칙이 생성되었다. 학력이 대제이상인 응답자들은 지역의 대기 환경오염도에 대하여 부정적인 시각을 가지고 있는 것으로 나타났다. 학력이 대제이상이면서 주관적 사회계층이 중하류층 이하, 결혼유무가 기혼, 연령이 평균이하인 응답자들은 지역의 대기 환경오염도에 대하여 부정적인 시각을 가지고 있는 것으로 나타났다.

5) 지역의 토양 환경오염도

지역의 토양 환경오염도의 연관성 규칙 결과는 <표 6>과 같다.

<표 6> 지역의 토양 환경오염도 연관성 규칙 적용 결과

규칙	지지도	신뢰도	전항값 1	전항값 2
1	10.42	82.54	학력 = 대제이상	주관적 사회계층 = 중하류층 이하
2	11.83	82.46	학력 = 대제이상	결혼유무 = 기혼

지역의 토양 환경오염도에 대한 연관성 규칙 적용 결과 총 2개의 규칙이 생성되었다. 학력이 대제이상이면서 주관적 사회계층이 중하류층 이하, 결혼유무가 기혼인 응답자들은 지역의 토양 환경오염도에 대하여 부정적인 시각을 가지고 있는 것으로 나타났다.

5. 결론

현재 경상남도는 경상남도 도민들을 대상으로 매년 환경, 교통 등의 부문에 대하여 사회지표 조사를 실시하고 있다. 그러나 3년 주기로 매년 설문 문항을 다르게 하여 사회지표 조사를 실시하고 있어 도민들의 환경의식에 대한 분석 시 연도별로 각각 분석을 실시해야 함으로써 유기적인 분석이 가능하지 못한 실정이며, 특정 연도의 사회지표 조사 자료에서는 환경의식 분석에 사용할 환경 관련 문항들이 기타 연도에 비하여 작아 다양한 분석을 실시하지 못하고 있다. 그러므로 도민들의 환경의식에 대한 분석 시 연도별로 각각 분석을 실시해야 함으로서 유기적인 분석이 가능하지 못하여 분석의 한계점이 있다. 만일 각 연도의 사회지표 조사 자료를 결합하여 하나의 데이터 파일을 만들 수 있다면, 유기적인 분석이 가능하여 고부가 가치의 정보를 획득할 수 있을 것이다.

이에 본 논문에서는 경상남도에서 2001년과 2002년에 조사된 사회지표 조사 자료의 환경자료에 대하여 도민들의 환경 의식을 더욱더 총체적으로 분석하기 위하여 데이터 퓨전 기법을 이용하여 연관성 규칙을 적용하였다. 본 논문에서는 환경 자료 중 공통으로 가지는 변수에 개인 식별 가능한 변수가 없기 때문에 통계적 결합방법을 사용하여 퓨전을 실시하였으며, 데이터퓨전기법을 이용한 환경자료의 연관성 규칙 적용 결과 도민들의 환경의식을 더욱더 통합적으로 분석할 수 있었다. 향후 환경자료뿐만 아니라 다양한 분야의 자료에 대하여 데이터퓨전기법을 적용하여 효율적 데이터 사용과 데이터 분석의 질을 높일 수 있을 것으로 사료된다.

참고 문헌

1. 김호종, 김은주, 김명원 (2003), 신경망 기반 추천 모델의 성능향상을 위한 정보의 융합, *한국정보과학회 학술발표논문집*, Vol. 2003, No. 2483, pp.422-424.
2. 박성원, 권지웅, 최진영 (2001), 데이터 퓨전을 이용한 얼굴영상 인식 및

인증에 관한 연구, *퍼지 및 지능시스템학회 논문집*, Vol. 11, No. 4, pp.302-306.
3. 박희창, 조광현 (2005), 의사결정나무 기법을 이용한 사회지표조사 자료 분석, *The Journal of Korean Data Analysis Society*, Vol. 7, No. 3, pp.773-783.
4. 손소영, 이성호 (2000), 데이터 융합, 앙상블과 클러스터링을 이용한 교통사고 심각도 분류분석, *대한산업공학회/한국경영과학회 2000년 춘계공동학술대회 논문집*, Vol. 2000, pp.597-600.
5. 신형원, 손소영 (2000), 다구찌 디자인을 이용한 데이터 퓨전 및 군집분석 분류 성능 비교, *대한산업공학회/한국경영과학회 2000년 춘계공동학술대회 논문집*, Vol. 2000, pp.601-604.
6. 최기주, 정연식 (1998), 링크 통행시간 추정을 위한 데이터 퓨전 알고리즘의 개발, *대한교통학회지*, Vol. 16, No. 2, pp.177-195.
7. Agrawal R., Imielinski R., Swami A.(1993), Mining association rules between sets of items in large databases, *In Proc. of the ACM SIGMOD Conference on Management of Data*, Washington, D.C.
8. Agrawal R., Srikant R.(1994), Fast algorithms for mining association rules, *In Proc. of the 20th VLDB Conference*, Santiago, Chile.
9. Cheung D.W., Han J., Ng V., Fu A.W., Fu Y.(1996), A Fast distribution algorithm for mining association rules, *Int's Conf. on Parallel and Distributes Information System*, Miami Beach, Florida.
10. National Statistics (2003), National Statistics Code of Practice Protocol on Data Matching. http://www.statistics.gov.uk/about/consultations/general_consultations /downloads/ Protocol_ on_Data_Matching.pdf
11. Park J.S., Chen M.S., and Philip S.Y.(1995), An effective hash-based algorithms for mining association rules, *In Proc. of ACM SIGMOD Conference on Management of Data*, Washington, D.C.
12. Saygin Y., Vassilios S.V., Clifton C.(2002), Using Unknowns to Prevent Discovery of Association Rules, *2002 Conference on Research Issues in Data Engineering*.
13. Sergey B., Rajeev M., Jeffrey D.U., Shalom T.(1997), Dynamic itemset counting and implication rules for market data, *In Proceedings of ACM SIGMOD Conference on Management of Data*. Washington, D.C.
14. Toivonen H.(1996), Sampling Large Database for Association Rules, *Proceedings of the 22nd VLDB Conference Mumbai(Bombay)*, India.

[2007년 1월 접수, 2007년 4월 채택]