데이타마이닝을 이용한 지능적 이미지 검색 시스템 설계*

○
이충우, 나연묵
단국대학교 컴퓨터공학과

Design of Intelligent Image Retrieval System using Data Mining

○
Chung-Woo Lee, Yunmook Nah
Dept. of Computer Engineering, Dankook University

요약

본 논문은 이미지 검색 시스템의 사용자 인터페이스 모듈로부터 기록되는 사용자의 질의 로그 파일에 데이타마이닝 기법을 적용하여 후에 동일한 패턴의 질의시 보다 빠른 검색을 할 수 있는 방법을 제안한다. 사용자 인터페이스 모듈에서 사용자의 질의를 통해 얻어진 이미지 클래스를 질의 로그 파일에 기록하여 연관 규칙을 탐사하게 된다. 이때 찾아진 연관 규칙을 데이타베이스에 저장하여 새로운 검색 질의가 들어올 때 동일한 패턴을 찾아 검색이 예상되어지는 이미지 클래스로부터 검색을 함으로써 검색범위를 줄여 속도를 개선할 수 있고, 사용자의 질의 횟수를 줄일 수 있어 보다 효율적인 검색을 할 수 있다.

1. 서론

최근 들어 컴퓨터 처리 성능 향상과 저장 장치의 발달, 그리고 고해상도 카메라와 스캐너 등의 이미지 수집을 위한 하드웨어의 급증에 따라 이미지 데이타의 활용도가 점차 증가하고 있다. 이미지 데이타들은 개별적으로 웹 홈페이지에 올려져 있는 형태일 수도 있고, 전자 미술관/전자 박물관 등의 이미지 전문 데이타베이스에 대단위로 저장되어 있을 수도 있다. 이러한 이미지들의 양은 대단히 방대하므로 사용자가 원하는 이미지를 찾을 때, 이미지 데이타베이스에서 제공하는 방법을 이용하거나 웹 검색 엔진의 도움을 받게 된다.

웹 검색 엔진이나 이미지 데이타베이스에서 제공하는 검색 방법은 크게 두 가지로 나눌 수 있다. 첫째는 이미지와 관련된 텍스트, 키워드 등을 이용한 텍스트 기반의 검색[1, 2]이며, 둘째는 이미지 자체의 특성인 색상, 질감, 객체의 모양 등을 이용한 이미지 내용 기반 검색[3, 4]이다. 그러나 현재의 방법들은 과거의 검색 결과에 대한 사용자의 만족도를 고려하지 않고 있으며, 통합적으로 이미지 특성을 이용하지 못하여 검색 효율이 떨어진다는 단점을 지니고 있다. 그러므로 이를 보완하는 보다 지능적인 검색 기법이 요구된다.

따라서, 본 논문에서는 기존 이미지 검색 방법에 이미지 마이닝의 분류 기법을 이용하여 이미지를 분류하고 사용자의 질의 로그 파일에서 연관 규칙을 찾아 동일한 패턴의 검색시 보다 효율적인 검색을 지원할 수 있는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 데이타마이닝의 분류화(classification), 연관규칙

*이 논문은 한국과학재단의 특정 기초 연구비의 지원에 의한 것임.

(association rule) 탐사 방법을 소개하고 3장에서 내용기반 검색 모듈, 이미지 분류 마이닝 모듈, 사용자 인터페이스 모듈, 질의 로그 마이닝 모듈로 구성된 지능적 이미지 검색 시스템의 구조를 설명한다. 그리고 4장에서 결론을 맺는다.

2. 데이타마이닝

데이타마이닝은 대량의 데이타로부터 알려지지 않은 유용한 정보를 추출하여 의사 결정에 적용하는 과정[7, 8]이다. 그간 제안된 다양한 데이타마이닝 기법들에는 대용량 데이타베이스에 존재하는 여러 유용한 지식을 추출하는 방법으로서 분류화(classification), 군집화(clustering), 연관규칙 탐사(association rule), 경향분석(trend analysis), 패턴분석(pattern analysis) 등이 있다. 본 논문에서는 내용기반 검색 모듈로부터 추출되는 이미지의 특징 벡터를 데이타마이닝의 분류 기법을 적용하여 이미지를 분류하고 사용자 인터페이스로부터 기록되는 질의 로그 파일에서 연관 규칙을 찾는다.

2.1 분류화(classification) 기법

분류의 목적은 입력 데이타를 분석하여 각 클래스에 대해 정확한 표현(description)이나 모델을 개발하는 것이다. 여기서 입력 데이타는 속성이나 특성에 대한 레코드(예를 들어, training set)들로 구성된다. 클래스 분류는 테스트 데이타를 분류하기 위해 학습 데이타로부터 도출될 수 있다. 일반적으로 수백만 표본을 가진 큰 데이타 집합을 분류하기 위해서 의사 결정 트리 분류자(Decision Tree Classifier)를 사용한다[9]. 의사 결정 트리 분류자는 다른 분류 방법에 비해 상대적으로 빠르며 데이타베이스를 액세스할 수 있는 SQL 질의로 전환할 수 있다. 그림 1은 6개의 표본에 대한 의사 결정 트리 분류의 예이다.

예를 들면, B는 Age<=35인 경우는 Salary<=40을 만족하고, Age>35인 경우는 Salary<=50을 만족하는

Age	Salary	Class
30	65	G
23	15	B
40	75	G
55	40	B
55	100	G
45	60	G

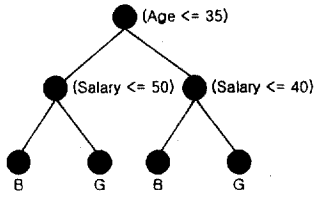

그림 1 의사 결정 트리의 예

클래스이다.

자료분류의 일반화된 형태는 속성간 종속성으로 표현할 수 있다. 애트리뷰트 종속성은 애트리뷰트 A1, A2, ..., Am이 주어졌을 때 f(A1, A2, ..., Am, 상수의 집합)→g(A1, A2, ..., Am, 상수의 집합)로 표현할 수 있다. 예를 들면 즉 "A1=c1이고 A2=c2이면 A3=c3이고 A4=c4이다"는 속성간 종속성을 생각할 수 있다. 하지만 f나 g는 임의의 값으로 주어질 수 있는 함수의 가지 수가 이론적으로 너무 많기 때문에 다루기 쉬운 문제가 아니다. 그래서 실제 도메인에 적용가능 하도록 f, g를 제한해야 한다. 예를 들면 f는 단순 술어들의 집합으로, g는 클래스의 레이블의 형태로 제약하게 되면 앞서 언급한 자료 분류 문제로 귀착된다.

2.2 연관 규칙(Association rule) 탐사 기법

트랜잭션의 모임으로부터 한 항목집합의 존재와 다른 항목집합과의 연관 관계를 요약할 수 있다. 연관관계는 항목들 사이에 존재하는 유사성 또는 패턴을 의미한다. 연관 규칙의 기본 개념을 설명하면, $I=\{i_1, i_2, ..., i_k\}$라는 항목(item)집합 즉 트랜잭션의 집합을 고려한다. 이 경우 $i_1, i_2, ..., i_k \subset I$이다. ϕ이 아닌 항목집합 X, Y에 대해 $X \subset I$, $Y \subset I$에 대한 연관 규칙 $X \rightarrow Y$는 $X \cap Y = \phi$의 특성을 갖는다. X는 규칙의 가정, Y는 규칙의 결과라고 한다. 항목 집합 I의 부분집합 X에 대해, $X \subseteq T$이면 T는 X를 만족한다고 정의한다. 최소지지도(minimum support threshold)를 만족하는 $X \subseteq I$를 Large 항목집합이라 한다[10].

생성된 연관규칙이 트랜잭션들의 상황을 얼마만큼 잘 뒷받침해 주는가는 두 가지의 척도로서 측정한다.

- 지지도(support degree): 생성된 연관규칙이 전체 아이템에서 차지하는 비율을 말한다. 즉 데이타베이스에 속한 전체 트랜잭션의 개수 중 그 연관규칙을 지지하는 트랜잭션의 비율을 의미한다.
- 신뢰도(confidence degree): 연관규칙의 강도를 의미하며 전제부를 만족하는 트랜잭션이 결론부까지를 만족하는 비율을 말한다.

그림 2는 연관규칙 탐사과정의 예를 보인 것이다. 이는 Apriori 알고리즘에 의거한다[10]. 연관규칙 탐사과정은 크게 두 단계로 구성되는데, 첫단계로 높은 지지도를 갖는 아이템의 집합을 식별하는 작업이

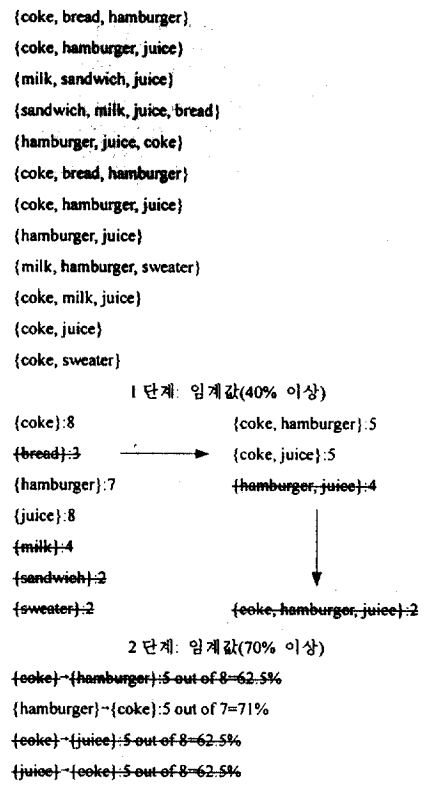

그림 2 Apriori 알고리즘을 이용한 연관규칙 탐사과정

며 두 번째 단계에서는 높은 신뢰도를 갖는 연관규칙을 도출하는 작업이다. 첫 단계를 수행하기 위해 데이타베이스의 트랜잭션을 조회하여 항목별 빈도수를 구한 다음 최소한의 지지도를 만족하는 항목만을 고른다. 이때 지지도에 대한 임계값이 40%라 주어졌다면 이를 만족하는 아이템들의 집합은 {coke}:8, {juice}:8, {hamburger}:7이 된다. 다음으로 이 항목들의 조합으로 구성된 사건항목 집합에 대해 최소 지지도 이상을 만족하는 항목들을 반복하여 찾는다면 {coke, hamburger}:5와 {coke, juice}:5의 결과를 얻는다. 두 번째 단계로 신뢰도의 임계값이 70%라 하였을 때 생성된 연관규칙은 최종적으로 {hamburger}{coke}:5/7=71%을 얻게된다.

3. 데이타마이닝을 이용한 지능적 이미지 검색 시스템 설계

본 논문에서 제안하는 데이타마이닝을 이용한 지능적 이미지 검색 시스템의 전반적인 구조는 그림 3과 같다. 웹에 흩어져 있는 이미지들은 내용 기반 검색

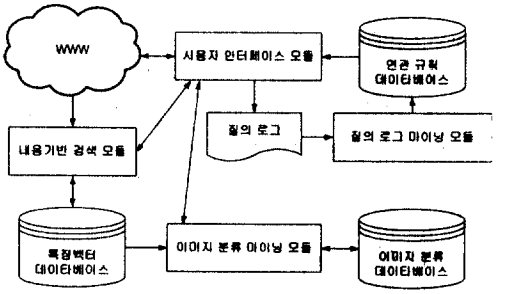

그림 3 데이타마이닝을 이용한 지능적 이미지 검색 시스템

모듈을 통해 텍스트, 키워드, 색상, 질감, 객체의 모양 등의 이미지 특징벡터 데이타베이스가 형성된다[3]. 이 특징벡터 레코드들을 이미지 분류 마이닝 모듈의 입력 데이타로 받아 각 특징표현(feature description)에 대해서 이미지들을 분류하여 저장하게 된다. 사용자 인터페이스 모듈에서는 사용자로부터 검색 질의를 입력받아 내용 기반 검색 모듈을 통해 검색 결과를 사용자에게 보여준다. 그러는동안 사용자의 질의 내용이 질의 로그(Query Log) 파일에 기록되고 질의 로그 마이닝 모듈에서 이것을 분석해 연관 규칙을 탐사한다. 추출된 연관 규칙이 연관 규칙 데이타베이스에 저장되고 이것을 통해 후에 동일한 패턴의 질의가 들어오면 연관 규칙에 따라 이후 검색되어질 대상들에 대한 검색이 함께 이루어져 사용자에게 보여준다.

3.1 내용 기반 검색 모듈

웹에 흩어져 있는 이미지들을 검색하기 위해서는 웹 문서를 자동으로 분석, 인덱싱하고 이미지를 주제별로 분류하는 웹 에이전트를 이용하여 웹에서 이미지를 수집해야한다. 이미지의 목록화와 검색을 지원하기 위해 텍스트와 시각 정보를 동시에 이용한다.

이미지 데이타베이스에 대한 검색은 이미지 자체보다는 이미지의 색상, 모양, 질감 등을 기술하는 특징벡터(feature vector)를 이용하여 질의를 기술한 형태로 가장 가까운 이미지 데이타를 검색하는 기법이 많이 사용되고 있다. 또는 키워드 형태의 설명(description) 데이타에 대한 텍스트 정보 검색 기법도 많이 사용되고 있다.

이미지 데이타베이스 검색 시스템은 키워드를 통해 이미지 카탈로그(image catalog)를 검색하거나 축약 이미지(thumbnail)를 이용해 브라우징하는 방법을

사용하고 있다.

IBM의 QBIC(Query By Image Content)[3]의 경우에는 색상, 질감, 객체의 모양 등의 특성을 이용해 이미지를 검색한다. QBIC에서 특징벡터는 점(point)으로 표현되고, R*-트리를 이용해 인덱싱되어 저장된다. 스케치 이미지는 축약이미지(thumbnail) 형태로 저장된다.

Chabot[4]이 지원하는 개념(concept) 질의란 직접적인 색상의 비교보다 상위 수준의 내용 검색 질의로 'sunset', 'snow'와 같은 문맥 정보를 활용하는 질의이다.

질의는 시각적 예제(visual example)를 이용해 제시한다. 각 저장 이미지에 대해 그래픽 특징 벡터를 추출한다. 이 특징 벡터는 흑색 픽셀의 공간 분포, 공간적 빈도 등을 나타낸다. 검색은 그래픽 특징 벡터간의 유사성 척도(similarity measures)를 기반으로 진행된다.

본 시스템에서는 웹으로부터의 이미지들의 내용검색을 위해 텍스트, 키워드, 색상, 질감, 객체의 모양 등의 특징 벡터들을 추출하고 이것을 특징벡터 데이타베이스에 저장하여 QBE(Query By Example) 형태로 제시되어지는 질의 이미지의 질의 특징 표현과의 유사도 연산을 하게 된다.

다음은 주로 사용되는 특징 표현(feature description)들이다.

- 이미지의 칼라 히스토그램
- 이미지의 객체의 모양
- 이미지의 질감
- 이미지의 키워드

3.2 이미지 분류 마이닝 모듈

내용 기반 검색 모듈에 의해 추출된 이미지의 특징 표현(feature description)은 벡터 집합 형태로 생성된다. 이미지 분류 마이닝 모듈은 주어진 의사 결정 트리에서 특징표현(feature description) 분류자로 이미지를 각 이미지 클래스로 분류하게 된다[11]. 이때 하나의 이미지는 각 특징 속성에 따라 이미지클래스에 속하게 된다. 즉, 하나의 이미지는 특징 속성 개수만큼의 이미지클래스에 속한다. 이는 이미지가 검색될 때 하나의 속성에 따른 검색이 다른 어떤 속성과 연관성 있게 검색이 되는지 알아보고자 하기 때문이다.

그림 4에서 이미지1은 의사 결정 트리의 특징표현

이미지번호	특징 속성	특징 벡터 값
1	ColorHiF	000c0009 0000007c 7e141300 ....
1	TextureF	00000000 00000265 0000168f 000002b9
1	DrawFeaF	04d7d744 04d856f5 04040408 ....
1	TextFeaF	Sunflower
2	ColorHiF	00050000 09000080 00000000 ....
2	TextureF	00000000 0000023a 00001809 00000201
2	DrawFeaF	15040434 35323300 d7f04126 ....
2	TextFeaF	sweet gum tree leaf liquidambar leaf

(a) 특징벡터 데이타베이스

이미지클래스	이미지번호
1	1, 8, 38, 58
2	2, 36, 48
3	3, 14, 67,84

(b) 이미지 분류 데이타베이스

그림 4 이미지 특징벡터로부터 이미지 분류

분류자에 의해서 ColorHiF에 대해서 이미지클래스1에 속하게 되었다. TextureF에 대해서는 또 다른 이미지클래스에 속하게 될 것이다. 이미지클래스1의 경우는 이미지1의 ColorHiF에 대한 이미지클래스였다. 따라서 이미지클래스1에 속해 있는 또 다른 이미지 8, 38, 58 역시 특징 속성 ColorHiF에 대해 유사한 특징 벡터 값을 갖는 이미지들일 것이다.

3.3 사용자 인터페이스 모듈

사용자 인터페이스 모듈은 사용자의 질의를 입력받아 이미지 분류 마이닝 모듈을 통해 이미지 클래스를 찾아낸 후, 연관 규칙 데이타베이스에서 동일한 패턴을 찾는다. 질의 패턴을 찾지 못했을 경우 질의 내용은 {<시간>, <사용자ID>, <질의 이미지 클래스>}의 형태로 질의 로그(Query Log) 파일에 기록된다. 이때 사용자ID는 사용자 PC의 IP주소가 된다. 질의 로그 파일은 후에 연관 규칙을 찾는데 사용된다. 질의에 대한 결과는 내용기반 검색 모듈을 통해 사용자에게 보여준다. 연관 규칙 데이타베이스에서

```
f = a query image feature attribute;
v = a query image feature vector;

q = find the class by f, v from image class DB;
r = search a rule includes q from a RuleBase;
if (exist r) then do begin
        rc = a result part of r;
        show images are retrieved by f, v on rc;
        end
 else do begin
        write q into a Log;
        show images are retrieved by f, v;
        end
```

그림 5 질의 패턴 검사 알고리즘

사용자의 질의 패턴이 발견되면 연관 규칙의 결과부분의 이미지 클래스를 포함한 범위에서 질의가 이루어져 사용자에게 검색 결과를 보여준다. 그림 5는 질의 패턴을 검사해 사용자에게 검색 결과를 보여주는 알고리즘이다.

질의 로그(Query Log)의 요소는 다음과 같다.

- 시간: 트랜잭션이 일어난 시간
- 사용자ID: 사용자PC의 IP주소를 갖는 환경변수값
- 질의 이미지클래스: 의사 결정트리로 분류된 질의 이미지의 이미지클래스

예를 들어, 99년 4월 1일 10시 2분 50초에 111.111.111.111의 IP를 갖는 PC로부터 사용자가 A이미지의 Color Histogram이 유사한 이미지를 검색하고자 할 때, 우선 A이미지의 Color Histogram을 구하고(A이미지의 Color Histogram은 '000c0009 0000007c 7e14 1300 ...'), Color Histogram '000c0009 0000007c 7e141 300 ...'가 속하는 이미지 클래스를 찾는다(Color Histogram '000c0009 0000007c 7e141300 ...'가 속하는 이미지 클래스는 10). 따라서 질의 로그는 다음과 같이 생성된다.

99-04-01 10:02:50, 111.111.111.111, 10

3.4 질의 로그 마이닝 모듈

사용자 인터페이스 모듈로부터 생성된 질의 로그(Query Log) 파일을 마이닝하여 연관 규칙을 탐사한다. 우선 그림 6의 (a)에서 (b)와 같이 Log 파일을 트랜잭션 항목집합으로 변환해야 하는데, 트랜잭션 항목집합은 한 사용자가 원하는 이미지를 찾기 위한 질의 과정의 나열이다. 즉, 하나의 트랜잭션 항목집합에는 첫째, 동일한 사용자ID를 가진 트랜잭션들이어야 하고, 둘째, 트랜잭션들의 시간 간격이 주어진 최소 시간 간격(minimum time gap) 이하이어야 한다. 사용자가 검색을 시작해서 마칠 때까지는 연속적인 검색이 이루어지고 검색을 마치고 새로운 검색을 하기까지는 많은 시간 간격이 있다고 가정한다. 그림 7은 질의 로그 파일로부터 트랜잭션 항목집합을 추출하는 알고리즘이다.

```
99-04-01 10:02:50,111.111.111.111,10
99-04-01 10:04:04,111.111.111.111,5
99-04-01 10:04:07,222.222.222.222,24
99-04-01 10:06:02,111.111.111.111,8
99-04-01 10:07:20,222.222.222.222,32
99-04-01 10:09:40,333.333.333.333,19
```

(a) 질의 로그 파일

TID	Items	
1	{10, 5, 8}	111.111.111.111 의 트랜잭션
2	{24, 32}	222.222.222.222 의 트랜잭션
3	{19, 38, 29, 56}	333.333.333.333 의 트랜잭션

(b) 트랜잭션 항목집합

전제부	결과부	
{10}	{8}	{10}→{8}
{19, 29}	{56}	{19, 29}→{56}

(c) 연관 규칙 데이타베이스

그림 6 질의 로그 파일로부터 연관 규칙 탐사

```
// D[tid] is a tid's itemset
// T[tid] is a last transaction time of tid's itemset
// U[tid] is a trasaction userid of tid's itemset

g = a minimum time gap;
tid = 1;
forall transaction tr∈Log file do begin
  u = a userid part of tr;
  t = a time part of tr;
  d = a query image class part of tr;
  k = 0;
  while (k++<tid) {
        if (U[k]=t) then
          if (|T[k]-t|<g) then do begin
                insert d into D[k];
                T[k] = t;
                break;
                end
        }
  if (k=tid) then do begin
        U[tid] = u;
        T[tid] = t;
        D[tid] = d;
        tid++;
        end
end
```

그림 7 트랜잭션 항목집합 추출 알고리즘

이렇게 추출된 트랜잭션 항목집합으로부터 트랜잭션의 연관 규칙을 탐사하게 된다[10]. 그림 6의 (c)와 같이 찾아진 연관 규칙은 연관 규칙 데이타베이스에 저장되어 사용자의 새로운 질의시 질의 패턴을 검사하게 되는 대상이 된다.

4. 결론

본 논문은 과거 검색 결과에 대한 사용자의 만족도를 고려하지 않고 있는 기존 내용 기반 검색 시스템을 개선하여 이미지 마이닝의 분류 기법을 이용해서 이미지를 분류하고 사용자의 질의 로그 파일에서 연관 규칙을 탐사하여 동일 패턴 검색시 보다 효율적인 검색을 지원할 수 있는 방법을 제안하였다.

본 논문에서 제안한 시스템을 통해 이미지를 검색시 과거 사용자의 이미지에 대한 검색 패턴 정보를 활용해 질의 이미지와 연관된 검색이 예상되어지는 이미지 클래스에서 검색함으로써 검색 범위를 줄여 속도를 개선할 수 있고, 사용자의 질의 횟수를 줄일 수 있어 보다 효율적인 이미지 내용 검색을 수행할 수 있다.

내용기반 검색 모듈을 통해 생성된 텍스트, 키워드, 색상, 질감, 객체의 모양 등의 특징벡터 데이타베이스는 이미지 분류 마이닝 모듈을 통해 이미지 클래스들로 분류된다. 사용자 인터페이스 모듈로부터 입력받은 질의는 질의 로그 파일에 기록되고, 이 질의 로그 파일은 질의 로그 마이닝 모듈을 통해 연관규칙이 탐사된다. 추출된 연관규칙을 통해 동일한 패턴의 질의가 들어오면 연관 규칙에 따라 이후 검색되어질 대상들에 대한 검색이 함께 이루어진다.

이미지 분류 마이닝과정에서 이미지를 분류하는데는 이미지 도메인 지식을 필요로 한다. 따라서 향후 본 시스템의 구현시 적절한 이미지 도메인을 설정하고 그 이미지 도메인 지식을 적용하여 이미지 분류 마이닝의 의사결정 트리 분류자를 설계할 예정이다.

참고문헌

[1] Lycos Inc., http://www.lycos.com

[2] Infoseek Corporation, http://www.infoseek.com

[3] M. Flickner, et al., "Query By Image and Video Content: The QBIC System," IEEE Computer, Vol.28, No.9, IEEE CS & IEEE, pp.23-32, Sept. 1995.

[4] V. E. Ogle and M. Stonebraker, "Chabot: Retrieval from a Relational Database of Images," IEEE Computer, Vol.28, No.9, IEEE CS & IEEE, pp.40-48, Sept. 1995.

[5] 김정자, 이도헌, "데이타 마이닝 기술 및 연구동향", 정보과학회, 제16권, 제9호, 정보과학회지, pp.6-14, 1998년 9월.

[6] 윤종필, 김희숙, 최옥주, "데이타 마이닝의 유용성", 정보과학회, 제16권, 제9호, 정보과학회지, pp.15-23, 1998년 9월.

[7] R. Agrawal, T. Imielinski and A. Swami, "Database Mining: A Performance Perspective," IEEE Transformation on Knowledge and Data Engineering, pp. 914-925, 1993.

[8] M. S. Chen, J. Han, and P. S. Yu, "Data Mining: An Overview from Database Perspective," IEEE Transactions on Knowledge and Data Engineering, Vol.8, No. 6, pp.866-883, Dec. 1996.

[9] M. Mehta, R. Agrawal, and J. Rissanen, "SLIQ:A Fast Scalable Classifier for Data Mining," EDBT, 1996.

[10] R. Agrawal, T. Imielinske, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD, pp.207-216, 1993.

[11] Osmar R. Zaiane, Jiawei Han, Ze-Nian Li, Jean Hou, "Mining Multimedia Data," Proc. CASCON'98: Meeting of Minds, Toronto, Canada, November 1998.