

지능적 이미지 검색 시스템을 위한 질의 패턴 탐사*

이충우, 나연묵
단국대학교 컴퓨터공학과

Mining Query Patterns for the Intelligent Image Retrieval System

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

요 약

본 논문은 지능적 이미지 검색 시스템을 위한 질의 패턴 탐사를 제안한다. 지능적 이미지 검색 시스템은 이미지 검색시 질의 로그로부터 사용자의 검색 패턴을 탐사하여 패턴에 따라 연관된 검색을 동시에 수행함으로써 검색 효율을 높일 수 있는 시스템이다[1]. 본 논문은 이 시스템의 질의 로그 마이닝 과정에서 필요한 질의 패턴 탐사 방법을 제안한다. 연관 규칙의 경우 단편적인 연관 관계만이 탐사되기 때문에 사용자의 질의 패턴으로 사용하기에 효과적이지 못하다. 따라서 본 논문에서는 연관 규칙을 개선하여 하나의 항목과 연관된 패턴을 표현할 수 있는 형태를 제안한다. 질의 패턴을 사용하면 사용자의 패턴을 탐사하는 응용에서 좀 더 효율적으로 사용할 수 있다.

1. 서론

컴퓨터 시스템의 발달과 데이터베이스 시스템의 사용 증가로 컴퓨터에 저장되는 데이터의 양은 폭발적으로 증가하고 있다. 데이터마이닝(data mining)은 대용량의 데이터에서 숨겨진 유용한 패턴을 추출하는 방법[3]이다. 데이터에서 숨겨진 패턴을 탐사하는 방법 중 연관 규칙 탐사[4]가 가장 많이 사용되고 있고, 그 응용이 많다.

연관 규칙 탐사는 초기 대용량의 운영계 데이터베이스로부터 의사 결정을 도울 수 있도록 데이터를 분석하고 레포팅하는 일로 시작되었으나, 현재 기존 검색 시스템의 과거 사용자 검색에 대한 만족도를 반영하기 위해 데이터마이닝을 통합하는 연구[1]와 같이 데이터마이닝 통합환경 시스템을 구축하려는 연구들이 활발히 이루어지고 있다.

한편, 연관 규칙에는 단편적인 규칙만이 탐사되기 때문에 위와 같이 사용자의 패턴을 탐사하여 사용하는 응용에는 부적절하다. 따라서 본 논문에서는 연관 규칙의 특성을 알아보고, 패턴을 표현할 수 있는 형태로 개선하여 질의 패턴 탐사 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 지능적 이미지 검색 시스템을 소개하고, 3장에서 데이터마이닝의 연

관 규칙 탐사 기법과 그 특성을 살펴본다. 4장에서 지능적 이미지 검색 시스템을 위한 질의 패턴 탐사 방법을 제안하고, 5장에서 결론을 맺는다.

2. 지능적 이미지 검색 시스템

지능적 이미지 검색 시스템은 기존 내용 기반 이미지 검색 시스템에 과거 사용자의 검색 결과에 대한 만족도 고려와 통합적인 이미지 특성을 이용하여 검색 효율을 높인 시스템[1]으로 그 구조는 그림 1과 같다.

데이터베이스에 저장된 이미지들은 내용기반 검색 모듈을 통해 텍스트, 키워드, 색상, 질감, 객체의 모양 등의 이미지 특징 벡터를 추출하여 특징벡터 데이터베이스에

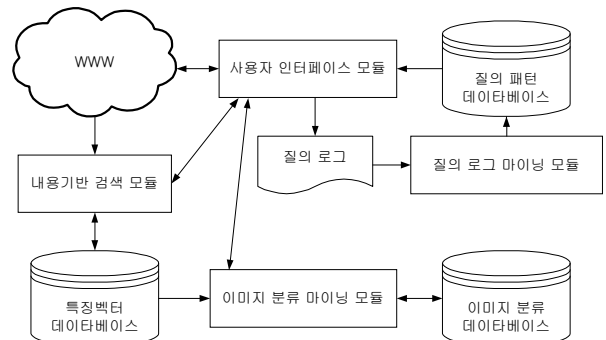


그림 1 데이터마이닝을 이용한 지능적 이미지 검색 시스템

*이 논문은 한국과학재단의 특정 기초 연구비의 지원에 의한 것임.

저장하고, 이 특징 벡터를 이미지 분류 마이닝 모듈에서 개념 계층(concept hierarchy)를 사용하여 분류하게 된다. 사용자가 이 시스템에서 이미지를 검색시 사용자의 질의 내용이 사용자 인터페이스 모듈로부터 질의 로그(query log)로 기록되고, 질의 로그 마이닝 모듈로부터 질의 패턴을 탐사하여 저장하여 다음 사용자의 검색시 도움을 주게 된다.

3. 연관 규칙 탐사 기법

연관 규칙은 한 항목들의 그룹과 다른 항목들의 그룹 사이에 강한 연관성이 있음을 밝혀준다. 이러한 연관성은 새로운 패턴을 찾아낼 수 있다.

정의1 연관 규칙(Association Rule)

$$R: X \rightarrow Y (X, Y \subseteq I, X \cap Y = \emptyset)$$

“만일 한 트랜잭션이 X를 지지한다면, 또한 어떤 확률에 의해 Y도 지지할 것이다”

X: 규칙의 조건부(antecedent)

Y: 규칙의 결과부(consequent)

지지도(support) = $P(X \cup Y) / N(D)$

신뢰도(confidence) = $\text{supp}(X \cup Y) / \text{supp}(X)$

3.1 연관 규칙 탐사 과정

연관 규칙 탐사 과정은 크게 두 단계로 구성되는데, 첫 번째 단계로 높은 지지도를 갖는 아이템의 집합(빈발 항목집합)을 식별하는 작업이며 두 번째 단계는 높은 신뢰도를 갖는 연관 규칙을 도출하는 작업이다.

첫 번째, 빈발 항목집합을 발견하는 단계는 데이터베이스의 트랜잭션을 조회하여 항목별 빈도수를 구한 다음 최소한의 지지도를 만족하는 항목만을 고른다. 다음은 빈발 항목집합의 특성들[2]이다.

특성1 부분집합의 지지도

만일 $A \subseteq B$ 이면, $\text{supp}(A) \geq \text{supp}(B)$ 이다.

특성2 빈발하지 않은 집합들의 상위집합들은 빈발하지 않다.

만일 $\text{supp}(A) < s_{\min}$ 이라면, 특성1에 의하여 $\text{supp}(B) \leq \text{supp}(A) < s_{\min}$ 이기 때문에 A의 모든 상위집합 B는 빈발하지 않을 것이다.

특성3 빈발 항목집합들의 부분집합들은 빈발하다.

만일 $\text{supp}(B) \geq s_{\min}$ 이라면, 특성1에 의하여 $\text{supp}(A) \geq \text{supp}(B) \geq s_{\min}$ 이므로 B의 모든 부분집합 A는 D에서 또한 빈발할 것이다. 특히, 만약 $A = \{i_1, i_2, \dots, i_k\}$ 가 빈발하다면, 그것의 모든 k개의 (k-1)-부분집합들도 빈발하다.

두 번째, 연관 규칙 탐사 단계는 빈발 항목집합의 조합으로부터 최소한의 신뢰도를 만족하였을 때 연관 규칙

이 생성된다. 다음은 문제점으로 대두되는 연관 규칙의 특성들이다.

특성4 연관 규칙의 중복성

만일 빈발 항목집합{A, B, C}가 있다고 하면, 빈발 항목집합의 특성3에 의해서 이 빈발 항목집합의 부분집합들도 빈발하다. 그러므로 연관 규칙 $A \rightarrow B$, $A \rightarrow C$ 와 같이 A에 연관된 규칙이 중복하여 생성될 수 있다.

특성5 연관 규칙의 분리성

빈발 항목집합{A, B, C}에 대해서 $A \rightarrow B$, $B \rightarrow C$ 와 같은 연관 규칙이 생성될 수 있다. 이것은 $A \rightarrow B \rightarrow C$ 와 같은 연관성이 분리되어 표현된다.

3.2 연관 규칙 탐사의 응용

연관 규칙 탐사는 응용성이 아주 높아 많은 연구가 이루어지고 있다. 기본적인 연관 규칙을 응용하여 각 항목의 분류를 포함하는 연관성을 찾아내는 일반화된 연관 규칙 탐사, 시간의 변이를 추가한 순차 패턴, 분산된 정보를 접근하는 패턴을 탐사하는 순회 패턴, 주기적인 연관성 등에 연구가 이루어져 이를 활용하는 소프트웨어 개발에도 연구가 집중되고 있다.

4. 질의 패턴 탐사

연관 규칙은 위에서 설명된 문제점들로 인해 지능적 이미지 검색 시스템과 같은 응용에서 사용자의 검색 패턴 탐사로 사용하기에 부적절하다. 따라서 연관 규칙의 중복성을 제거하여 조건부에 제시된 항목과 연관된 모든 항목을 결과부에 표시한다면 응용에서 패턴을 사용하는 데 효과적이다.

정의2 질의 패턴(Query Pattern)

$$P: X \rightarrow Y (X, Y \subseteq I, X \cap Y = \emptyset)$$

“만일 한 트랜잭션이 X를 지지한다면, 또한 어떤 확률에 의해 Y와 같은 패턴도 지지할 것이다”

X: 패턴의 조건부(antecedent)

패턴을 유발하는 하나의 항목으로 표현

Y: 패턴의 결과부(consequent)

조건부 항목과 연관된 패턴으로 항목들의 집합으로 표현

이 질의 패턴의 특성은 다음과 같다.

특성6 하위 패턴의 신뢰도

만약 $A \rightarrow B$, C와 같은 질의 패턴이 있을 경우, $\text{conf}(P) = \text{supp}(A \cup B \cup C) / \text{supp}(A) \geq c_{\min}$ 이다. 빈발 항목집합 특성1에 의해서 {A, B, C}의 부분집합

{A, B}에 대해서 $\text{supp}(A \cup B) \geq \text{supp}(A \cup B \cup C)$ 가 만족한다. 이것은 $\text{supp}(A \cup B)/\text{supp}(A) \geq \text{supp}(A \cup B \cup C)/\text{supp}(A) \geq c_{\min}$ 이므로, $A \rightarrow B$ 와 같은 질의 패턴도 만족한다. 마찬가지로 $A \rightarrow C$ 도 성립한다.

특성7 하위 패턴의 지지도

만약 $A \rightarrow B$, C 와 같은 질의 패턴의 경우, $\text{conf}(P) = \text{supp}(A \cup B \cup C)/\text{supp}(A) < c_{\min}$ 일 경우, $\text{supp}(A \cup B \cup C) \leq \text{supp}(A \cup B)$ 이다. 만약 $\text{supp}(A \cup B \cup C) = \text{supp}(A \cup B)$ 이라면 $\text{supp}(A \cup B \cup C)/\text{supp}(A) = \text{supp}(A \cup B)/\text{supp}(A) < c_{\min}$ 이므로 하위 패턴 $A \rightarrow B$ 도 최소 신뢰도를 만족하지 못한다. $\text{supp}(A \cup B \cup C) < \text{supp}(A \cup B)$ 이라면 $\text{supp}(A \cup B \cup C)/\text{supp}(A) < \text{supp}(A \cup B)/\text{supp}(A)$ 이므로 이 경우에 만 최소 신뢰도를 만족하는지 살펴보아야 한다.

4.1 질의 패턴 탐사 과정

질의 패턴 탐사 과정도 연관 규칙 탐사와 마찬가지로 두 단계의 과정을 거치게 된다. 첫 번째 단계는 빈발 항목집합의 발견이다. 이 단계는 연관 규칙과 같은 방법을 사용한다. 두 번째 단계는 질의 패턴 탐사 과정이다. 질의 패턴 탐사 단계는 최대 빈발 항목집합으로부터 하나의 조건부 항목으로부터 나머지 항목들이 패턴으로 성립되는지 최소 신뢰도를 기준으로 탐사하게 된다. 질의 패턴 탐사 알고리즘은 그림 2와 같다. 그림 3은 연관 규칙 탐사 과정과 패턴 탐사 과정을 비교한 것이다.

5. 결론

본 논문에서는 지능적 이미지 검색 시스템에서 사용자의 질의 패턴을 탐사하기 위한 방법을 제안하였다. 질의

```

procedure discovery(antecedent, itemset)
  if ( $\text{supp}(\text{itemset})/\text{supp}(\text{antecedent}) \geq c_{\min}$ )
    then insert a query pattern;
    else for each (k-1)-subset of consequent do
      begin
        // 질의 패턴 특성 7
        if ( $\text{supp}(\text{itemset}) < \text{supp}((k-1)\text{-subset})$ )
          then discovery(1-subset, (k-1)-subset);
        end
      end procedure
for each itemset from large itemset do
  begin
    for each 1-subset of itemset do
      begin
        discovery(1-subset, itemset);
      end
    // 질의 패턴 특성 6
    delete all subset from large itemset;
  end

```

그림 2 질의 패턴 탐사 알고리즘

최소 지지도 = 0.4

빈발 항목집합	지지도
A	0.5
B	0.7
C	0.6
D	0.4
E	0.8
A, B	0.5
A, C	0.5
B, C	0.6
D, E	0.4
A, B, C	0.5

질의 패턴 탐사

최소 신뢰도 = 100%

{A, B, C} $A \rightarrow B, C$ (0.5/0.5) $B \rightarrow A, C$ (0.5/0.7) $C \rightarrow A, B$ (0.5/0.6)
 $B \rightarrow A$ (0.5/0.7) $C \rightarrow A$ (0.5/0.6)
 $B \rightarrow C$ (0.6/0.7) $C \rightarrow B$ (0.6/0.6)

{D, E} $D \rightarrow E$ (0.4/0.4) $E \rightarrow D$ (0.4/0.8)

연관 규칙 탐사

최소 신뢰도 = 100%

$A \rightarrow B$ (0.5/0.5) $A, B \rightarrow C$ (0.5/0.5)
 $A \rightarrow C$ (0.5/0.5) $A, C \rightarrow B$ (0.5/0.5)
 $B \rightarrow C$ (0.6/0.7) $B, C \rightarrow A$ (0.5/0.6)
 $B \rightarrow A$ (0.5/0.7)
 $C \rightarrow A$ (0.5/0.6)
 $C \rightarrow B$ (0.6/0.6)
 $D \rightarrow E$ (0.4/0.4)
 $E \rightarrow D$ (0.4/0.8)

그림 3 질의 패턴 탐사와 연관 규칙 탐사

패턴을 탐사하기 위해 연관 규칙 탐사 기법을 살펴보고, 패턴 탐사에 사용시 대두되는 문제점을 제시하였다. 이를 해결하기 위해 본 논문은 연관 규칙의 표현 형태를 개선하여 질의 패턴을 제안하였으며, 연관 규칙과의 관계를 설명하였다. 그리고, 질의 패턴 탐사 알고리즘을 보였다.

향후 지능적 이미지 검색 시스템에 질의 패턴 탐사 기법을 적용하여 리콜(recall)과 정확도(precision) 측정을 통해 유용성을 검증할 예정이다.

참고문헌

- [1] 이충우, 나연목, “데이타마이닝을 이용한 지능적 이미지 검색 시스템 설계”, 한국멀티미디어학회 춘계학술 발표 논문집, 제2권, 제1호, pp.115-120, 1999년 5월.
- [2] 박종수, 유원경, 홍기형, “연관 규칙 탐사와 그 응용”, 정보과학회지, 제16권, 제9호, pp.37-44, 1998년 9월.
- [3] R. Agrawal, T. Imielinski, and A. Swami, "Data Mining: A Performance Perspective," *IEEE Transformation on Knowledge and Data Engineering*, pp. 914-925, 1993.
- [4] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *Proc. ACM SIGMOD*, pp.207-216, May 1993.