내용 기반의 멀티미디어 데이터 연관규칙 마이닝에 대한 연구

김 진 옥[†] · 황 대 준[††]

요 약

컴퓨터 처리기술과 저장기술 그리고 인터넷 등의 영향으로 멀티미디어 데이터의 양은 급속하게 증가하지만 체계적인 멀티미디어 데이터간의 연관규칙을 마이닝하는 연구는 초기 단계이다. 본 논문은 이미지 프로세싱 분야 및 내용기반 이미지 검색에 대한 기존 연구를 바탕으로 대형 영상 데이터 저장소에 저장된 이미지 데이터에서 재생성되는 항목간의 연관규칙을 찾으며 공간적 관계로 내용기반의 연관규칙을 마이닝하는 알고리즘을 제안한다. 제안된 연관규칙 탐색 알고리즘은 이미지의 색상, 질감, 모양 등 내용기반의 영상속성을 오브젝트 항목으로 하고 오브젝트가 이미지에서 재생성될 때를 이용하여 이미지간의 연관규칙을 찾고 오브젝트들이 이미지에서 차지하고 있는 공간적 위치관계를 통해 드러나지 않는 이미지간의 연관규칙을 마이닝한다. 본 논문의 재생성 항목을 고려한 연관규칙 알고리즘은 Apriori 알고리즘보다 빈번한 항목 집합을 찾아내는데 더 높은 성능을 갖는다는 것을 실험을 통하여 보여준다. 제안된 알고리즘은 동일한 정보원으로부터 받은 멀티미디어 데이터간의 연관성을 탐색하는데 특히 효과적이며 다양한 관련 응용분야에 적용할 수 있다.

A Study on Data Association-Rules Mining of Content-Based Multimedia

Jin Ok Kim[†] · Dae Joon Hwang[††]

ABSTRACT

Few studies have been systematically pursued on a multimedia data mining in despite of the overwhelming amounts of multimedia data by the development of computer capacity, storage technology and Internet. Based on the preliminary image processing and content-based image retrieval technology, this paper presents the methods for discovering association rules from recurrent items with spatial relationships in huge data repositories. Furthermore, multimedia mining algorithm is proposed to find implicit association rules among objects of which content-based descriptors such as color, texture, shape and etc. are recurrent and of which descriptors have spatial relationships. The algorithm with recurrent items in images shows high efficiency to find set of frequent items as compared to the Apriori algorithm. The multimedia association-rules algorithm is specially effective when the collection of images is homogeneous and it can be applied to many multimedia-related application fields.

키워드 : CBIR(Content-Based Image Retrieval), 멀티미디어(Multimedia), 데이터 마이닝(Data Mining)

1. 서 론

멀티미디어 데이터 마이닝은 멀티미디어 데이터 베이스에 저장된 드러나지 않는 패턴을 찾아내거나 멀티미디어 데이터간의 관계를 규명한다. 이미지나 비디오 데이터로부터 관련 속성을 추출해 내는 이미지 프로세싱 기술과 추출된 속성을 이용하여 데이터 베이스에 저장되어 있는 타 이미지와 비교한 후 필요로 하는 이미지를 검색하는 내용기반 이미지 검색 기술 그리고 속성 정보를 통해 데이터 저장소에 저장된 데이터간의 연관관계 등의 규칙정보를 찾아내고 분류하여 사용자 인터페이스로 보여주는 기술은 상호 관련이 있는 멀티미디어 데이터 마이닝의 복합 연구 분야이다

Fayyad[1, 2]의 위성 사진으로부터 패턴을 발견하는 방법을 기점으로 하여 멀티미디어 데이터를 추출하고 인덱싱하여 저장하고 보여주는 많은 기술들이 제안되었다. Czyzewski[3]는 오디오 데이터를 분석하고 오래된 레코딩에서 잡음을 제거하는 방법을 제안했으며 Chien [4]은 대형 이미지 데이터 베이스에서 이미지 프로세싱을 지원하는 지식기반의 인공지능 기술을 선보였다. Bhandari[5]는 멀티미디어 자원과 데이터 마이닝 응용분야를 결합했다.

† 정 회 원 : 성균관대학교 대학원 전기전자 및 컴퓨터공학부
†† 정 회 원 : 성균관대학교 전기전자 및 컴퓨터공학부 교수
논문접수 : 2001년 11월 28일, 심사완료 : 2002년 1월 4일

멀티미디어 데이터 마이닝은 영상적 데이터 속성을 포함한 데이터로부터 OLAP(Online analytical processing)과 OLAM (Online analytical mining)한 것이었다. OLAP와 OLAM을 위해 이미지와 비디오 속성을 프로세싱하거나 추출하는데 주로 멀티미디어 오브젝트의 크기, 데이터의 분포도, 키워드 등의 속성을 이용하여 데이터간의 특징과 연관성 및 분류 규칙을 마이닝한다. 그러나 이미지에서 색상, 질감, 형태 등 내용 기반의 속성과 오브젝트간의 공간위치와 시간변화에 따른 오브젝트 움직임 등 공간 관계에 따른 연관성을 이용하여 멀티미디어 데이터간의 연관규칙을 마이닝하는 연구방법은 초기 단계이다.

연관규칙 마이닝은 최근 데이터 마이닝 연구영역에서 광범위하게 연구되어 왔으며[1, 2, 9-11, 14, 17] 많은 알고리즘과 접근방법이 대형 데이터 베이스에서 여러 종류의 연관규칙 마이닝이 제안되었지만 관련 데이터 베이스는 문자와 숫자 그리고 트랜잭션 기반의 데이터였다. 몇몇 제안 알고리즘은 프로세스할 수 있는 형태로 변환한 영상 데이터 분야에 적용될 수 있지만 영상 데이타가 가진 이미지 정보의 특이성을 찾아내기는 어렵다. 어떤 영상 속성은 한 이미지에서 여러 번 발생할 수 있고 속성의 반복적 발생은 속성 그 자체보다 더 많은 정보를 제공할 수 있으며 이미지 데이터 오브젝트간의 공간적 관계를 통해서 찾아낸 데이터간의 연관규칙 역시 중요한 지식을 형성할 수 있기 때문에 기존 제안된 연관규칙의 적용은 이미지와 비디오 데이터로부터 연관규칙을 마이닝하는데 한계가 있다.

본 논문에서는 이미지 멀티미디어 데이터에서 내용기반의 영상속성을 오브젝트로 하여 자주 발생하는 오브젝트와 이 오브젝트간의 공간 관계를 이용하여 멀티미디어 연관규칙을 발견하는 방법과 이를 위해 재생성되는 항목과 항목간의 공간 관계를 통해 멀티미디어 연관규칙을 마이닝하는데 유효한 알고리즘을 제안한다.

2. 멀티미디어 연관규칙

본 논문의 목적은 이미지의 영상 속성과 이미지에 포함된 오브젝트간의 공간 관계를 이용해 데이터간의 연관규칙을 찾아내는 것이므로 비디오나 영상 데이타에 CBIR(Content-Based Image Retrieval)[13, 16, 18] 기술을 적용하여 오브젝트 기반의 이미지 분할 기법으로 이미지를 동일한 속성을 지닌 영역으로 분할한 후 저장한 이미지 데이터 베이스에서 연관규칙을 마이닝하는 방법을 찾는다. CBIR에서 이미지 분할은 이미지를 서로 연결되지 않는 영역으로 나누는 과정을 말한다. 영역은 특정 속성을 공유하는 픽셀의 집합으로 구성된다. 분할 알고리즘은 (i) 영역이 서로 연결된 상태에서 (ii) 영역은 해체되며 ($R_i \cap R_j = 0$, for $i \neq j$) (iii) 분할은 어떤 픽셀이 특정 영역에 할당되면 완료되고 모든 영역의 합집합은 전체 이미지가 됨 ($\bigcup_{k=1}^{m} R_k = I$)을 가정한다.

2.1 이미지 재생성 항목의 연관규칙

이미지는 영상 속성인 어떤 항목의 트랜잭션으로 모델링되고 이미지 ID는 트랜잭션 ID로 모델링된다. 본 논문에서는 이미지를 트랜잭션으로 표현하지만 오브젝트가 이미지에서 반복적으로 발생함을 고려하여 지지도는 이미지의 수 대신 이미지의 오브젝트 수를 반영한다. Apriori 알고리즘[1]과 같은 기존 연관규칙에서는 k 항목을 가진 후보 집합인 C_k를 구성할때 반복되는 항목에 대해 논의가 없으나 본 논문에서는 반복되는 항목을 고려한 연관규칙을 정의하고 이를 재생성 항목의 연관규칙이라 한다.

정의 1. 재생성 항목의 연관규칙은 이미지와 비디오 프레임에서 영상 오브젝트 속성의 연관규칙을 말하며 다음과 같은 형태이다

$$\alpha S_1 \wedge \beta S_2 \wedge \cdots \wedge \gamma S_n \rightarrow \delta T_1 \wedge \lambda T_2 \cdots \wedge \mu T_m (C\%)$$

여기서 C는 연관규칙의 신뢰도이며, S_i, T_j는 위상, 영상, 움직임 또는 다른 이미지 설명자에서 파생한 속성이고, i, j는 속성자의 하위 속성을 말한다. $\alpha, \beta, \gamma, \delta, \lambda, \mu$는 오브젝트 속성 또는 항목의 발생 빈도를 표시하는 정수 계수이다. S가 a 발생 빈도를 갖는다면 aS는 참이다. 연관규칙에서 S_i, T_j는 이미지 크기, 비디오 지속시간, 관련 키워드와 같은 다른 속성도 될 수 있다.

정의 2. 이미지 집합 D에서 $\sigma(S/D)$로 표시되는 속성 S의 지지도는 주어진 개념적 수준에서 S에 대한 D의 이미지 오브젝트의 백분율로 트랜잭션 기반의 지지도와는 달리 오브젝트 기반의 지지도이며 이미지가 주어진 속성을 갖고 있는 백분율을 의미한다. σ는 지지도 자체를 표시한다. 멀티미디어 연관규칙 $S \rightarrow T$의 신뢰도는 $\sigma(S \wedge T/D)$대 $\sigma(S/D)$의 백분율이며 S를 확인하는 D상의 이미지 오브젝트로 T가 확인되는 가능성이다.

정의 3. 패턴 p는 최소 지지도가 σ'이고 최대 지지도가 Σ'일때 $\sigma' < \sigma(p) < \Sigma'$이면 이미지 집합 D에서 충분히 빈번하다.

정의 4. 이미지 집합 D에서 멀티미디어 연관규칙 $S \rightarrow T$는 S와 T가 충분히 빈번하게 발생하고 $S \rightarrow T$의 신뢰도가 최소 신뢰도 값보다 크다면 D에서 충분히 강하다.

연관규칙의 강함과 최소 지지도 σ'와 최대 지지도 Σ'의 값은 이미지의 내용기반 속성이 적용된 수준에 달려있다. 색상, 질감, 움직임, 방향과 같은 속성은 계층 개념구조로 정의된다. 이용자가 선택하는 속성에 대한 개념 수준과 이미지의 해상도 수준에 따라 σ'와 Σ'는 더 높아지거나 낮아진다. 이미지 I에서 항목 수만큼의 트랜잭션과 이미지를 동일한 속성별로 분할한 영역 수가 주어지면 두 종류의 멀티미디어 연관규칙인 재생성 영상속성을 가진 내용기반의 규칙과 재생성 공간 관계를 가진 규칙을 고려할 수 있다.

멀티미디어 연관규칙이 이용하는 영상 속성은 색상이나 질감과 같이 계층개념에 의해 정의한 오브젝트의 속성과 이미지의 분할된 영역사이의 위상(수직관계 : V-next-to, 수평관계 : H-next-to, 중첩 : Overlap, 포함 : Include)이다. 각 속성 S는 두 오브젝트 O_a와 O_b간의 중첩(O_a, O_b)과 같은 관계를 설명한다. 각 오브젝트는 다차원이고 관련 바이너리 속성은 한 개 이상의 관계에 대해 조인(join) 내용을 포함하며 공간 속성은 같은 오브젝트 값에서 재발생할 수 있다.

2.2 오브젝트를 포함한 테스트 이미지 생성

본 논문에서 제안된 알고리즘을 설명하고 그 효과를 측정하기 위해 먼저 랜덤한 속성과 분할된 이미지를 이용하여 오브젝트를 가진 테스트 이미지들을 생성한다. 그 과정은 다음과 같은 단계를 거친다 ; (i) 랜덤한 배경색상을 가진 각 이미지 n개를 생성한다. (ii) 각 이미지는 분할된 k개의 영역을 갖고 있다. (iii) 각 영역에서 색상, 픽셀수, 질감, 모양, 위치 등의 속성을 랜덤하게 생성한다. 주어진 영역에 대해 새로운 위치를 가진 랜덤 프레임을 생성함으로써 각 영역에 방향성을 부여한다. 그 결과 이미지를 포함하고 있는 프레임들의 n개 집합 분류가 생성된다. 각 프레임은 m개의 이미지 갯수를 가지고 있고 각 이미지는 각각 k개의 오브젝트 갯수를 가지고 있다. 다른 프레임 집합의 이미지는 다른 오브젝트 갯수를 가지고 있다. 각 이미지 프레임의 이미지 식별번호를 가진 $I_1, I_2, \cdots, I_n$에서 오브젝트들은 $O_{(1,1)}$ $O_{(1,2)}, O_{(2,1)} \cdots O_{(n,a)}$과 같은 식별번호를 갖고 $Colour_1$ $Texture_1$, $Size_1$, $Shape_1$, $Direction_1$등과 같은 해당 오브젝트의 색상, 질감, 크기, 모양, 방향 등의 속성 벡터값을 취한다. 연관규칙 탐색 알고리즘은 이 속성 정보를 마이닝하여 재생성 항목이 있는 내용기반의 멀티미디어 연관규칙을 찾는다. 오브젝트의 공간 위상이 주어지면 I_1이미지에서 오브젝트 식별번호 $O_{(1,1)}$를 갖는 오브젝트는 속성 V-Next-to관계인 $\{O_{(1,3)}$, $O_{(1,5)}\}$, H-Next-to 관계인 $\{O_{(1,2)}\}$, Overlap 관계인$\{O_{(1,7)}\}$ Include 관계인 $\{O_{(1,9)}\}$등의 위치 속성값을 갖는다. $O_{(1,2)}$오브젝트는 별도의 위치 속성값을 갖는다. 이 위치 속성값은 재생성 공간 관계를 갖는 멀티미디어 연관규칙을 찾는데 이용한다.

3. 재생성 항목에 대한 멀티미디어 연관규칙 마이닝

재생성 항목을 가진 빈번한 항목 집합을 찾는 방법은 먼저 빈번한 한 개 항목 집합을 찾아서 이 항목이 이미지에서 얼마나 자주 다시 발생하는지 확인하고 최고 발생 빈도 항목을 찾을 때까지 반복해서 이용할 k 엘리먼트 집합에 통합한다. 빈번하지 않은 항목은 지지도 값을 적용하여 탈락시킨다. 만약 Apriori 알고리즘을 데이터 집합에서 빈번하게 발생하는 항목 집합을 발견하는데 적용한다면 한 이미지에서 재생성 항목을 갖는 모든 항목 집합은 무시할 가능성이 높다.

3.1 최고 발생도 탐색 알고리즘

재생성되는 영상 속성 오브젝트에 기반한 강한 멀티미디어 연관규칙을 찾는 방법에서는 단일 차원과 단일 해상도의 오브젝트 개념을 이용한다. 이미지의 트랜잭션은 오브젝트로 구성되고 오브젝트가 트랜잭션에서 계속 반복된다. 관련 알고리즘을 설명하기 위해 예 1을 설정한다.

〈표 1〉 이미지 트랜잭션 집합과 최고 발생 빈도 항목

(a) 이미지 트랜잭션 집합 D_1

이미지 ID	오브젝트
I_1	$\{O_2, O_2, O_4, O_5\}$
I_2	$\{O_2, O_3, O_3, O_4\}$
I_3	$\{O_1, O_2, O_3, O_3\}$
I_4	$\{O_2, O_2, O_2, O_3, O_3, O_4\}$
I_5	$\{O_5, O_5, O_6\}$
I_6	$\{O_7, O_8\}$

(b) 후보 집합 C_1과 발생 빈도 M

오브젝트	지지도	최고 발생 빈도
$\{O_1\}$	1	1
$\{O_2\}$	7	3
$\{O_3\}$	5	2
$\{O_4\}$	3	1
$\{O_5\}$	3	2
$\{O_6\}$	1	1
$\{O_7\}$	1	1
$\{O_8\}$	1	1

예 1 : <표 1>에서 제시한 이미지를 트랜잭션 집합 D_1으로 간주한다. 각 이미지는 반복되는 오브젝트의 집합이다. 오브젝트의 지지도를 결정하기 위한 첫 번째 데이터 베이스 스캔이 이루어지면 <표 1>의 오른쪽 값이 결정된다. 한 개 후보 항목 집합 C_1은 지지도와 함께 모든 단일 오브젝트를 표시하고 M은 한 이미지에서 오브젝트의 최고 발생

빈도 수를 표시한다. 단순화를 위해 지지도는 절대값으로 하여 최소 지지도 σ'은 2로 하고 최대 지지도 Σ'은 5로 한다. C_1에서 $\sigma'=2$보다 작은 지지도를 가진 k항목 집합을 필터링하여 탈락시킨다. 이 과정을 통해 <표 2>의 빈번한 한 개 항목 집합인 F_1이 생성된다. Σ'보다 지지도가 큰 빈번한 항목 집합은 다른 오브젝트와 결합하여 보다 지지도가 낮은 항목 집합을 생성할 수 있으므로 모든 빈번한 항목을 찾을 때까지 $O_2(\sigma(O_2/D_1)>\Sigma')$는 너무 빈번하게 나타나더라도 제거하지 않는다.

<표 2> 한 개 항목 집합과 필터링된 트랜젝션 집합

(a) 한 개 항목 집합 F_1과 M

오브젝트	지지도	최대 발생 빈도
$\{O_2\}$	7	3
$\{O_3\}$	5	2
$\{O_4\}$	3	1
$\{O_5\}$	3	2

(b) 필터링된 트랜젝션 집합 D_2

이미지 ID	오브젝트
I_1	$\{O_2, O_2, O_4, O_5\}$
I_2	$\{O_2, O_3, O_3, O_4\}$
I_3	$\{O_2, O_3, O_3\}$
I_4	$\{O_2, O_2, O_2, O_3, O_3, O_4\}$
I_5	$\{O_5, O_5\}$

F_1을 반영하여 D_1로부터 빈번한 오브젝트를 포함하지 않는 모든 트랜잭션을 제거함으로써 빈번한 오브젝트만을 가진 이미지 트랜잭션 집합인 D_2가 생성된다. <표 3>의 두 개 후보 항목 C_2의 생성은 F_1요소를 조인하여 이루어진다. M의 값은 Apriori 알고리즘과는 달리 이미지에서 오브젝트의 복제를 고려하여 트랜잭션에서 한 번 이상 발생한 동일한 오브젝트의 새로운 쌍을 생성하는데 이용한다. <표 3>에서 두 개 후보 항목 집합 $\{O_2, O_2\}$와 $\{O_3, O_3\}$는 이러한 방법으로 생성한 것이다. <표 4>의 세 개 후보 항목 집합 C_3는 F_2를 조인하여 생성되고 F_3은 C_3에서 최소 지지도 값이 2가 아닌 세 개 항목 집합을 제거함으로써 생성된다. M의 값은 역시 <표 4>의 $\{O_2, O_2, O_2\}$와 같은 항목 집합을 생성하기 위해 사용된다.

네 개 후보 항목 $\{O_2, O_3, O_3, O_4\}$집합은 빈번한 세 개 항목 집합을 조인하고 불필요한 것을 전지하는 방법으로 생성된다. <표 3>과 <표 4>의 경우처럼 $\{O_2, O_2, O_3, O_3\}$, $\{O_2, O_2, O_4, O_4\}$,$\{O_2, O_3, O_4, O_4\}$는 $\{O_2, O_2, O_3\}$과 $\{O_2, O_4, O_4\}$, $\{O_3, O_4, O_4\}$가 F_3에 없으므로 제거한다.

<표 3> 빈번한 두 개 항목 집합

(a) 후보 집합 C_2

오브젝트	지지도
$\{O_2, O_3\}$	3
$\{O_2, O_4\}$	2
$\{O_3, O_4\}$	2
$\{O_4, O_5\}$	1
$\{O_2, O_2\}$	2
$\{O_3, O_3\}$	2
$\{O_5, O_5\}$	1

(b) 두 개 항목 집합 F_2

오브젝트	지지도
$\{O_2, O_3\}$	3
$\{O_2, O_4\}$	2
$\{O_3, O_4\}$	2
$\{O_2, O_2\}$	2
$\{O_3, O_3\}$	2

<표 4> 빈번한 세 개 항목 집합

(a) 후보 집합 C_3

세 개 항목 집합	지지도
$\{O_2, O_3, O_4\}$	1
$\{O_2, O_2, O_3\}$	1
$\{O_2, O_2, O_4\}$	3
$\{O_2, O_3, O_3\}$	2
$\{O_3, O_3, O_4\}$	2
$\{O_2, O_2, O_2\}$	1

(b) 세 개 항목 집합 F_3

세 개 항목 집합	지지도
$\{O_2, O_2, O_4\}$	3
$\{O_2, O_3, O_3\}$	2
$\{O_3, O_3, O_4\}$	2

<표 5> 빈번한 네 개 항목 집합

(a) 후보 집합 C_4

네 개 후보 항목 집합	지지도
$\{O_2, O_3, O_3, O_4\}$	2

(b) 네 개 항목 집합 F_4

네 개 항목 집합	지지도
$\{O_2, O_3, O_3, O_4\}$	2

다섯 개 항목 집합이 더 이상 없다면 위 과정의 결과는 최대 지지도 Σ'보다 높은 지지도를 가진 항목 집합이 없는 모든 F_i이 된다. 다음은 이렇게 찾아낸 빈번한 k항목 집합이다.

$\{O_2, O_3, O_3, O_4\}, \{O_2, O_2, O_4\}, \{O_2, O_3, O_3\}, \{O_3, O_3, O_4\},$
$\{O_2, O_3\}, \{O_2, O_4\}, \{O_3, O_4\}, \{O_2, O_2\}, \{O_3, O_3\}$

충분히 빈번한 항목 집합이 주어지면 $0<k<T$를 조건으로 "$(k-T)$항목집합 → T항목집합" 형태의 k항목 집합으로부터 모든 규칙을 만들어 냄으로써 강한 연관규칙을 찾아낼 수 있다. 이 연관규칙은 주어진 신뢰도 임계값보다 더 높은 규칙 신뢰도를 가진다. 그 결과 100%의 신뢰도 임계값을 가진 다음과 같은 규칙을 유도한다.

(1) $\{O_3, O_4\} \rightarrow \{O_2, O_4\}\{100\%\}$

(2) $\{O_3, O_3, O_4\} \rightarrow \{O_2\}\{100\%\}$

(3) $\{O_3, O_3\} \rightarrow \{O_2\}\{100\%\}$

(4) $\{O_2, O_2\} \rightarrow \{O_4\}\{100\%\}$

(5) $\{O_3, O_3\} \rightarrow \{O_4\}\{100\%\}$

(6) $\{O_4\} \rightarrow \{O_3\}\{100\%\}$

(7) $\{O_4\} \rightarrow \{O_2\}\{100\%\}$

이 규칙들의 스캔을 통해 복제한 오브젝트를 고려하여 다음 규칙을 생성한다.

$O_3 \wedge O_4 \rightarrow O_2 \wedge O_4[100\%]$, $2O_3 \wedge 2O_4 \rightarrow O_2[100\%]$,
$2O_3 \rightarrow O_2[100\%]$, $2O_2 \rightarrow O_4[100\%]$, $2O_3 \rightarrow O_4[100\%]$
$2O_4 \rightarrow O_3[100\%]$, $\{O_4\} \rightarrow \{O_2\}\{100\%\}$

예 1과 관련한 내용은 다음의 내용기반 멀티미디어 연관규칙 마이닝에 대한 알고리즘으로 구현된다. 예 1에서 사용된 지지도는 단순화를 위한 절대값이다. k항목 집합을 위한 지지도는

$$\frac{D_k \text{에서 } k \text{ 항목 집합 수}}{\Sigma_{\forall \text{ transaction t}} \binom{|t|}{k}}$$

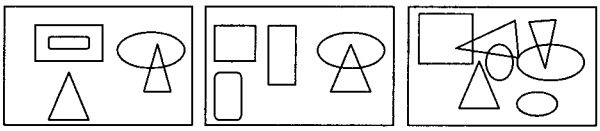
여기서 $\binom{|t|}{k}$는 t 트랜잭션에서 오브젝트들의 k 조합이다.

begin
(1) $C_1 \leftarrow$ {1개 후보 항목 집합과 그 지지도}
(2) $F_1 \leftarrow$ {자주 발생하는 1개 항목 집합과 그 지지도}
(3) $M \leftarrow$ {자주 발생하는 1개 항목 집합의 이미지에서의 최대 발생 정도}
(4) k 항목 집합의 카운트 수 ($total[1 \cdots k]$)
(5) while ($i \leftarrow 2$; $F_{(i-1)} \neq \phi$; $i \leftarrow i+1$) {
(6) $C_i \leftarrow (F_{i-1} \triangleright \triangleleft F_{i-1}) \bigcup \{y \oplus X \mid X \in F_{i-1} \wedge$ 카운트$(y, X) < (M[y]-1)\}$
(7) $C_i \leftarrow C_i - \{c \mid c$의 $(i-1)$ 항목 집합 $\notin F_{i-1}\}$
(8) $D_i \leftarrow FilterTable(D_{i-1}, F_{i-1})$
(9) while (각 이미지 I in D_i) {
(10) while (각 c in C_i) {
(11) $c.$ 지지도 $\leftarrow c.$ 지지도 + 카운트(c, I)
(12) }
(13) }
(14) $F_1 \leftarrow \left\{ c \in C_i \mid \frac{c.\text{ 지지도}}{\text{전체 } i \text{ 아이템집합}} > \sigma' \right\}$
(15) }
(16) Result $\leftarrow \bigcup_i \{c \in F_i \mid i > 1 \wedge c.$지지도$< \Sigma' \}$
end

(최고 발생도 탐색 알고리즘)

4. 공간 관계를 갖는 멀티미디어 연관규칙 마이닝

내용기반의 멀티미디어 연관규칙이 영상 속성을 사용하는데 반해 여기에 더해 공간 관계를 갖는 멀티미디어 연관규칙은 공간적 속성을 갖는 확장된 관계를 고려한다. 공간 관계를 갖는 멀티미디어 연관규칙은 먼저 공간 속성을 최소화한 후에 최고 발생도 탐색 알고리즘을 이용한다. 연관규칙의 정책은 빈번한 한 개 또는 두 개 항목을 찾고 이 항목에 공간적 속성을 통합하여 그 결과를 공간적 관계를 갖는 멀티미디어 연관규칙의 한 개 후보 항목 집합으로 한다. 그 다음 빈번한 공간 속성의 k항목 집합을 찾는데 최고 발생도 탐색 알고리즘을 이용한다. 공간 관계를 갖는 멀티미디어 연관규칙은 빈번한 공간 속성 $S(X, Y)$에 대해 X와 Y가 각각 빈번하게 발생해야 하며 두 개 항목 집합 $\{X, Y\}$가 빈번히 발생해야 함을 기반으로 한다. 공간 관계를 고려한 연관규칙은 다양한 공간 위상을 이용하므로 데이터 집합에 존재하지 않는 후보까지 포함해 대단히 큰 수의 공간 위상 항목 후보를 생성한다. 따라서 k항목 집합의 빈번한 공간 속성을 계산하기 전에 항목의 후보 수를 최소로 줄이는 것이 필요하기 때문에 본 논문에서는 일차원의 동일한 이미지 예를 분석한다.



(그림 1) 오브젝트를 가진 이미지의 예

예 2 : (그림 1)에서 일차원의 세 개 이미지를 고려해서 이미지 오브젝트 사이의 공간적 관계를 포함하는 연관규칙을 찾고자 한다. 이 내용은 연관한 개념 계층을 가진 색상, 질감 등과 같은 속성 차원에도 적용할 수 있다. 공간 관계를 갖는 강한 연관규칙을 찾아내는 것은 공간 속성의 빈번한 연관성을 찾는 것에 있다. 최소 지지도 임계 값 $\sigma' = 3$으로 할 때 이미지 집합의 첫 번째 스캔으로 세 개의 빈번한 항목을 찾아낸다.

〈표 6〉 공간위상을 갖는 빈번한 한 개 항목 오브젝트 집합

(a) 빈번한 오브젝트 쌍

오브젝트 쌍	{○,○}	{○,△}	{△,○}	{△,△}	{□,○}	{□,△}	{□,□}
지지도	1	3	1	2	3	3	1

(b) 빈번한 공간적 속성

한 개 항목 집합	Overlap {○,△}	H-Next -to {○,△}	H-Next -to {□,○}	H-Next -to {□,△}	H-Next -to {□,□}	V-Next -to {○,△}	V-Next -to {△,□}
지지도	3	1	3	3	1	1	2
최고 발생 빈도	2	1	2	2	1	1	1

○, △, □ 오브젝트는 세 가지 이미지에서 각각 나타나거

나 이미지에서 최대 두 번 나타난다. 이 세 가지 빈번한 오브젝트 항목을 고려하여 데이터 집합의 두 번째 스캔에서는 빈번하게 나타나는 항목 쌍을 찾아낸다. <표 6>의 (a)는 오브젝트 쌍의 지지도를 보여준다. 그 중 세 가지가 $\sigma' \geq 3$으로 충분히 빈번하며 공간적 속성과 쌍을 이룬다. ○, △, □ 오브젝트간에 네 가지 공간적 속성(H-next-to, V-next-to, overlap, include)이 존재한다면 이것으로 열두 가지 위상을 만들 수 있지만 (그림 1)의 데이터 집합의 스캔에서는 단지 일곱 가지 위상만을 볼 수 있다. 스캔 과정에서 항목의 지지도와 이미지에서의 최대 발생 빈도를 계산한다. <표 6>의 (b)는 스캔 결과 최고 발생도 탐색 알고리즘의 첫 번째 단계에서 찾아낸 빈번한 한 개 항목 집합이다. 최고 발생도 탐색 알고리즘은 그 다음 빈번한 k항목 집합을 찾아내는데 적용되어 Overlap(○, △), H-Next-to(□, △) ; Overlap(○, △), H-Next-to(□, ○) ; H-Next-to(□, △), H-Next-to(□, ○) ; Overlap(○, △), H-Next-to(□,△), H-Next-to(□, ○)과 같은 항목 집합을 찾아낸다.

여기서 파생한 연관규칙은 H-Next-to(□, ○)∧H-Next-to(□, △) → Overlap(○, △)[100%]이다.

예 2와 관련한 공간 관계를 가진 멀티미디어 연관규칙 마이닝은 공간 관계 탐색 알고리즘으로 구현한다.

begin
(1) $P_1 \leftarrow$ {빈번한 내용기반의 속성 항목}
(2) $P_2 \leftarrow \{P_1 \times P_1$에서 빈번한 쌍$\}$
(3) $C_1 \leftarrow \{P_2 \times${공간적 속성집합}과 그 지지도$\}$
(4) $F_1 \leftarrow \{C_1$에서 빈번한 1개 항목 집합$\}$
(5) 최고 발생도 탐색 알고리즘의 라인 3에서 라인 16까지 적용
end

(공간 관계 탐색 알고리즘)

5. 실험 및 결과분석

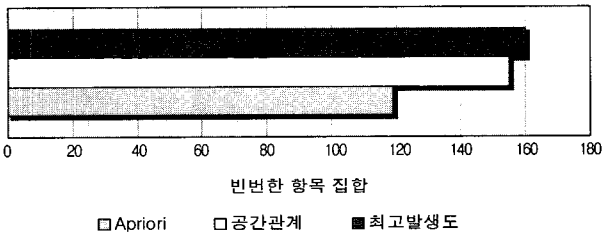
알고리즘의 확장성과 성능을 비교하기 위해 각 이미지당 15개 정도의 오브젝트를 가진 가상 이미지를 설정하고 다른 크기의 이미지 집합을 생성했다. 생성한 데이터 집합을 두 번 스캔한 후에 최고 발생도 탐색 알고리즘의 성능을 볼 수 있다.

Apriori 알고리즘은 트랜잭션 수 기반의 지지도를 이용하기 때문에 Apriori 알고리즘과 본 논문에서 제안한 알고리즘을 비교하기 위해 최고 발생도 탐색 알고리즘을 오브젝트 기반의 지지도 대신 이미지 트랜잭션 기반의 지지도를 가진 것으로 구현했다. 그러나 공간 관계 탐색 알고리즘은 정의 2에서 설명한 것처럼 오브젝트 기반의 지지도를 이용했다.

〈표 7〉 알고리즘별 다른 이미지 수에 대한 초당 평균 실행 시간

이미지 수	최고 발생도 탐색 알고리즘	공간관계 탐색 알고리즘	Apriori 알고리즘
1K	1.04	1.1	0.52
10K	11.34	11.84	5.31
20K	22.45	22.58	10.54
40K	44.73	45.02	21.02
60K	61.07	63.44	30.98
80K	86.24	88.23	41.93
100K	110.29	115.74	50.15

<표 7>은 세 가지 알고리즘에 대한 평균 실행 시간을 보여 준다. Apriori 알고리즘과 최고 발생도 탐색 알고리즘은 이미지 트랜잭션 기반의 지지도를 적용하여 $\sigma' = 0.05$로 설정하고 공간 관계 탐색 알고리즘은 오브젝트 기반의 지지도를 적용하여 $\sigma' = 0.0035$로 설정했다. 데이터 집합의 이미지 크기를 고려할 때 최고 발생도 알고리즘 및 공간 관계 탐색 알고리즘은 1초 동안 평균 1000개 이미지를 처리하고 있다. Apriori 알고리즘과 최고 발생도 탐색 알고리즘의 주요 차이점은 후보 항목 집합을 명확히 한 것과 이미지에서 재생성되는 항목의 발생을 계산한 것이다. 최고 발생도 탐색 알고리즘은 재생성되는 항목 집합을 잘 발견한다.



(그림 2) 알고리즘 별 빈번한 항목 집합 검색 결과

(그림 2)는 세 가지 알고리즘으로 찾은 빈번한 항목 집합의 평균 수를 보여준다. 최고 발생도 탐색 알고리즘은 160개의 다른 빈번한 k항목 집합을 찾고 공간 관계 탐색 알고리즘은 155개를 찾으며 Apriori 알고리즘은 119개를 찾는다. 최고 발생도 탐색 알고리즘과 공간 관계 탐색 알고리즘의 차이는 다른 지지도 값을 적용하였기 때문이다. Apriori 알고리즘에 비해 최고 발생도 탐색 알고리즘의 성능 저하 비용은 더 많은 빈번한 항목 집합을 찾음으로써 그리고 잠정적으로 재생성 항목을 가진 유용한 연관규칙을 찾음으로써 보상받을 수 있다.

6. 결 론

본 논문에서는 기존에 발표된 데이터 마이닝 연관규칙 알고리즘에서 고려하지 않은 재생성 항목을 기반으로 이미

지에서 재생성되는 오브젝트의 내용기반 영상 속성간에 존재하는 연관규칙을 찾아내기 위한 두 가지 알고리즘인 최고 발생도 탐색 알고리즘과 공간 관계 탐색 알고리즘을 제안했다. 최고 발생도 탐색 알고리즘은 색상, 모양, 크기, 질감 등의 속성 벡터 값을 이용하여 이미지에서 재생성되는 오브젝트 항목을 찾아 이 오브젝트간의 연관관계를 규명하며 공간 관계 탐색 알고리즘은 오브젝트가 이미지에서 차지하는 공간 위상을 이용하여 연관규칙을 발견한다.

실험 결과 제안된 두 개의 연관규칙 알고리즘은 이미지간의 연관규칙을 찾아내는데 있어서 Apriori 알고리즘보다 빠른 측정 결과를 보여주지는 못하지만 반복적이고 빈번한 재생성 오브젝트 항목을 찾는 데는 더 우수하기 때문에 같은 종류의 이미지가 모여 있는 저장소에서 이미지간의 연관 관계를 발견하는 멀티미디어 데이터 마이닝에 효과적이다. 마이닝의 대상이 되는 데이터가 동일한 영역을 찍은 감시 카메라의 영상, 특정 영역을 찍은 위성 사진, 환자의 CT 단층 촬용 사진처럼 같은 정보 제공 채널을 이용하거나 의미적으로 유사한 집합일 경우 더 강한 연관규칙을 찾아낼 수 있으므로 인물의 움직임, 영역의 변화, 환자의 신체적 이상여부를 감지하는데 적용 가능하다. 또한 비디오에서 개별 오브젝트를 정확하게 추출하여 확인하는 일은 매우 어려운 일이지만 본 논문에서 제안한 내용기반의 멀티미디어 데이터 마이닝 기술을 적용함으로써 오브젝트의 확인 가능성을 높일 수 있으며 공간 관계를 고려한 연관규칙은 이미지를 구성하는 영역의 움직임 벡터를 조건 필터로 하여 데이터 간 연관규칙 마이닝 시 높은 신뢰도를 보이므로 오브젝트의 일부인 비디오 시퀀스에서 같이 움직이는 이미지 영역을 찾아내는 데 사용될 수 있다.

참 고 문 헌

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining," AAAI/MIT Press, 1996.

[2] U. M. Fayyad, S. G. Djorgovski, and N. Weir. "Automating the analysis and cataloging of sky surveys," Advances in Knowledge Discovery and Data Mining, pp.471-493, AAAI/MIT Press, 1996.

[3] A. Czyzewski, "Mining knowledge in noisy audio data," In Proc. Second Int. Conf. on Knowledge Discovery and Data Mining, pp.220-225, 1996.

[4] I. Bhandari, E. Colet, J. Parker, Z. Pines and R. Pratap, "Advanced scout : Data mining and knowledge discovery in NBA data," Data Mining and Knowledge Discovery, Vol. 1, No.1, pp.121-125, 1997.

[5] S. Chien, F. Fisher, H. Mortensen, E. Lo, and R. Greeley, "Using artificial intelligence planning to automate science data analysis for large image databases," In Proc. Third Int. Conf. on Knowledge Discovery and Data Mining, pp.147-150, 1997.

[6] Stolorz, H. Nakamura, E. Mesrobian, R. Muntz, E. Shek, J. Santos, J. Yi, K. Ng, S. Chien, C. Mechoso, and J. Farrara, "Fast spatio-temporal data mining of large geographysical datasets," In Proc. Int. Conf. on KDD, pp.300-305, 1995.

[7] V. Tucakov and R. Ng, "Identifying unusual spatio-temporal trajectories from surveillance videos," In Proc. of 1998 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery(DMKD'98), Seattle, Washington, June, 1998.

[8] O. R. Zaiane, J. Han, Z. N. Li, J. Y. Chiang, and S. Chee, "MultimediaMiner : A system prototype for multimedia data mining," In Proc. ACM-SIGMOD, Seattle, 1998.

[9] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," In Proc. VLDB, pp.487-499, 1994.

[10] R. Miller and Y. Yang, "Association rules over interval data," In Proc. ACM-SIGMOD, pp.452-461, Tucson, 1997.

[11] R. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang, "Exploratory mining and pruning optimizations of constrained association rules," In Proc. ACM-SIGMOD, Seattle, 1998.

[12] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," In Proc. ACM-SIGMOD, Montreal, pp.1-12, 1996.

[13] J. R. Smith, C. S. Li, "Image classification and querying using composite region templates," Journal of CVIU, Academic Press, Vol.75, No.1-2, pp.165-175, 1999.

[14] T. Fukuda, Y. Morimoto, S. Morishita, and T.Tokuyama, "Data Mining using two dimensional optimized association rules : Scheme, algorithms, and visualization," In Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data, Montreal Canada, pp.13-23, June, 1996.

[15] 김진옥, 황대준, "다차원 데이터큐브를 이용한 멀티미디어 데이터 마이닝 연구", 한국정보과학회 추계학술대회논문집, 2001.

[16] M. J. Egenhofer and J. Sharma, "Topological relations between regions in γ^2 and z^2," In Advances in Spatial Databases (SSD '93), Singapore, 1993.

[17] Y. Fu and J. Han, "Meta-rule guided mining of association rules in relational databases," In Proc. 1st Int. Workshop Integration of Knowledge Discovery with Deductive and Object Oriented Databases, Singapore, pp.39-46, 1995.

[18] A. Natsev, R. Rastogi and K. Shim, "WALRUS : A similarity retrieval algorithm for image databases," In Proc. ACM-SIGMOD, Philadelphia, pp.395-406, 1999.

김 진 옥

e-mail : jinny@ece.skku.ac.kr

1985년 성균관대학교 문학사

1998년 성균관대학교 대학원 정보통신공학과(공학석사)

1998년~2000년 성균관대학교 전기전자 및 컴퓨터공학부 박사과정 수료

1992년~1994년 ㈜현대전자산업 정보통신사업본부 대리

1994년~1999년 ㈜현대정보기술 인터넷사업본부 과장

1999년~2000년 ㈜온세통신 온라인사업 팀장

2000년~2001년 ㈜유로코넷 기술담당 이사

관심분야 : Multimedia, Image Processing, Biometrics, Data mining, Recognition

황 대 준

e-mail : djhwang@skku.ac.kr

1978년 경북대학교 컴퓨터공학과(공학사)

1981년 서울대학교 컴퓨터과학과(이학석사)

1986년 서울대학교 컴퓨터과학과(이학박사)

1981년~1987년 한남대학교 전자계산학과 교수

1990년~1991년 미국 MIT 컴퓨터과학연구소 연구교수

1987년~현재 성균관대학교 전기전자 및 컴퓨터공학부 교수

관심분야 : 멀티미디어, 원격교육, 병렬처리, 가상교육, 지적재산권 보호 시스템