Multi-Dimensional Association Rule Mining in Multimedia Data

JinOk Kim, DaeJun Hwang
The School of Electrical & Computer Engineering,
SungKyunKwan University

요약

멀티미디어 데이터의 증가와 마이닝 기술의 발전으로 인해 멀티미디어 데이터에 대한 관심이 증가하고 있다. 본 논문에서는 특성특성화를 이용한 내용기반의 정보검색 기술과 다차원 데이터베이스 구축기술을 통해 멀티미디어 데이터에서 연관규칙을 찾아내는 멀티미디어 데이터마인트 시스템 프로토타입을 제안한다. 특히 멀티미디어 데이터의 확장성, 성능, 사용자 인터페이스, 그리고 연관규칙을 찾아내는 멀티미디어 데이터마인트 시스템의 구현방법에 대해 논한다. 필터링과 유사성 기반의 연관규칙을 찾아내는 멀티미디어 데이터마인트 시스템은 다양한 데이터의 연관규칙을 찾아내는 데 있어 경쟁력이 있다.

1. 서론

1.1 최근 인터넷과 웹기술의 발전 그리고 이를 기반으로 하는 멀티미디어 웹체크가 대량으로 증가하고 있다. 특히 웹기술의 발전과 이용자 수 증가에 따라 인터넷 데이터베이스를 추출하여 필요한 정보를 정확하게 찾는 것이 대두되고 있다[1]. 멀티미디어 데이터베이스는 멀티미디어 데이터베이스 개념의 확장된 패턴을 나타내며, 이는 멀티미디어 데이터와 관련된, 핵심적 지식의 추출을 이루어내는 분야로 데이터베이스의 한 영역이다. 현재 멀티미디어 데이터베이스는, 저장하고 인덱스하거나 검색하는 기능에 대한 연구가 다양하게 제안되고 있으므로 기존 멀티미디어데이터베이스와 데이터베이스 저장된 데이터베이스에 대한 연구를 포함하여 중요한 정보를 제공하고, 특히 멀티미디어 데이터베이스의 작동을 통해 멀티미디어 데이터마인드 시스템이 구현된다. 본 논문에서는 이미지데이터처리 과 데이터 마이닝 기술을 결합한 멀티미디어 기반 데이터마인드 시스템 프로토타입을 제안하고, 특히 멀티미디어 데이터베이스를 찾아내는 기술을 생성할 수 있는 방법을 제안한다. 이에 적용하는 방법으로 멀티미디어 데이터를 추출하여 데이터베이스화하는 과정에 효율적인 이미지 특성구조화 (feature localization)와 연관규칙 마이닝을 위한 다차원 데이터베이스 구축 및 검색알고리즘에 대해 설명한다.

2. 멀티미디어 데이터마인드 시스템 프로토타입

멀티미디어 데이터 마이닝은 컴퓨터비전, 컴퓨터 인공지능, 이미지프로세싱, HCK(Human Computer Interface) 등 이미지처리분야 기술과 마이닝 분야 기술을 종합하여 이루어내는 복합학문영역으로 이미지데이터를 제공하기로 약속한 내용기반의 이미지 추출시스템과 OLAP(On-Line Analytical Processing)기술, 그리고 데이터마인드의 특성화, 분류, 클러스터링, 연관관계등의 규칙을 찾아내어 일련의 데이터로부터 지식을 구축하는 마이닝기술을 근간으로 한다[2]. 본 논문에서 제안하는 멀티미디어 마이닝 시스템
은 이미지, 비디오 저장소로부터 필요한 멀티미디어 정보를 가져오는 검색에이전트와 이 검색에이전트가 가져온 멀티미디어 데이터를 사용하기 위한 컨텐츠 패러다임을 펼쳐놓고 변환하여 압축하는 전처리기모듈과 전처리기를 통해 의미있는 정보만을 추출하여 개별적인 속성을 가진 데이터요약과 정보추출내용으로 구성한 데이터베이스와 데이터베이스의 개별속성을 이용하여 구축한 멀티미디어 데이터클러스터에서 드릴다운(Drill-down), 롤업(Roll-up), 슬라이스(Slice), 다이스(Dice)와 같은 OLAP 분석 장치를 거쳐서 데이터간의 상호연관 관계를 찾아내고 분류하여 클러스터링한 후 예측하는 데이터마이닝 기술을 적용하는 분석모듈, 그리고 이용자가 원하는 지식을 탐구할 수 있도록 웹브라우저로 명확하게 보여주는 사용자 인터페이스 등으로 구성된다. 제안시스템은 그림 1과 같다.

그림 1. 멀티미디어 데이터마이닝 시스템의 구조

제안된 멀티미디어 데이터마이닝 시스템은 서 자재를 이미지와 비디오 데이터의 저장소로 본다. 이곳에서 사용량 멀티미디어 데이터를 추출하여 처리하는 내용기반 이미지추출 시스템을 이용하여 데이터 검색과 처리 그리고 데이터베이스 구축한다. 멀티미디어 데이터처리를 위해 적용되는 내용기반 이미지추출시스템은 이미지 추출에이전트, 이미지 내용 전처리기, 이용자 인터페이스, 데이터베이스에서 이미지, 비디오특성을 휘두며 매칭하는 검색커널으로 구성되어 있다.

3. 멀티미디어 데이터의 특성지식화 (Feature localization)

기존 내용기반 이미지 검색기술들에 쓰여지며 대부분의 기법은 색이나 질감과 같은 강한 이미지 특성에 의존한다. 이 기하적인 방법은 이미지 전체에서 추출된 단순한 도형자체 기반으로 사용되며 같은 형태, 같은 크기, 동일한 모양 등의 특성을 가진 이미지를 찾는 데 담당하는 데 많이 사용된다. 그러나 기하적인 방법은 대부분의 경우 대상체의 위치나 대상체의 특성(크기, 위치, 방향 등)을 식별하는 것이 어렵다. 보통 이미지에서 색과 질감 특성을 이용하여 내용검색을 시도하지만 이미지의 영상적 특성지식화(localization), 공간적 상관관계를 이용한 비디오영상의 속성을 이용한다. 따라서 본 논문에서는 전통적 이미지분할방법이 이에 대응하게 되는 비디오데이터에서에서 내용기반의 검색을 위해 적절한 이미지 전처리 방법이 있는 것을 표방하여 그에 대한 특성지식화에 대한 새로운 시도를 제안하고자 한다.

특성지식화는 [3]에서 연구한 이미지분할의 새로운 개념으로 이미지 분할은 이미지를 연결되지 않는 지역으로 나누는 것을 말한다. 이미지분할은 전체이미지를 분리된 영역으로 분할하는 과정이다. 이 영역은 또한 특성 즉 유사성을(회색 레벨 세기) 유사성각 등 여러 속성을 공유하는 펙셀 집합으로 구성된다. 전통적인 분할방법으로는 다음과 같은 내용을 포함하고 있다. [3]을 기준으로 R의 1개 영역이 대부분 결합되어 있으며 2개 영역은 분할된다. R \( \cap \) R = \( \emptyset \) \( \neq \) j (3) 문장은 어떤 펙셀이 어떤 영역에 할당되며 완료되고 여러 영역의 풍<ll>1의 이미지지식화를 나타낸다. 이에 이는 데이터에 있어서 각각 양 귀의 방법으로 펙셀을 이용한다.

특성지식화의 소재 (locale)는 이미지분할에 대해서 가장 작은 단위로 정의된 펙셀을 의미한다. 소재 (locale) \( L_x \)는 특성 x의 구리이다. \( L_x \)은 \( \epsilon \)의 지역성에 표현하는 타일함집합의 임의 \( L_x \)의 mass \( M(\epsilon) \) , (\( L_x \)에서 펙셀 수), centroid \( C(\epsilon) \) , (\( L_x \)에서 펙셀들의 중심), 펙셀 \( \sigma^2(\epsilon) \) , (\( L_x \)의 펙셀에서 중심까지의 Cartesain 펙셀, 소재의 모양 등 몇가지 기하학적 계수를 가지고 있다.

타일은 이상적인 정정단위이다. 타일은 펙셀의 중심은 펙셀의 중심과의 거리로 정의된다. 타일은 펙셀이 봉지거나 압축 펙셀을 모름을 볼 수 있고 펙셀의 영역도 있다. 펙셀이 이미지분할에 대한 정정 단위라라면 타일은 특성지식화에 대한 정정단위이다. 특성지식화는 중심을 가능한 가능한 단위로 할 수 있기 때문에 전장시가 불필요한 방법은 아니다. 특성지식화 방법에서는 이미지 간 위상관계 추정을 위해 소재(locale) 측위의 최소경계 원을 정의하여 이 원과 소재간의 관계를 비교하여 데이터로 검색할 수도 있다. 제안된 특성지식화에서는 근사화 위치가 된다. 이는 이미지와 비디오가 변화되는 내용기반의 이미지 검색 영역에서는 단순하고 명확한 위치 자체가 이루어질 가능성이 늘어지는 데. 대신, 여러 가지 특성과 그들의 공간적 관계와 관련된 프로세스.
가 더 명확하게 이루어진다. 그렇기 때문에 특정소역(local)은 추출하기 쉽고 이에 따라 특정국적화 방법이 필연에 의해 형성된 상세 영역보다 추출하기 쉽게 된다.

특성국적화를 통해 분할된 이미지는 각 계수값으로 메타데이터형태로 데이터베이스에 저장된다. 그리고 이용자의 취적에 따라 연관규칙에 의해 추출된다. 머신러닝 연관규칙은 데이터베이스에서 특정국적화계수와 같은 영상오브젝트의 특성을 연관하는 규칙을 가지고 있다.

\[ a P_1 \land \beta P_2 \land \ldots \land \gamma P_n \rightarrow \delta Q_1 \land \lambda Q_2 \land \ldots \land \mu Q_m \ (C \%) \]

여기서 C%는 규칙의 신뢰도이며, \( P_i, \ i \in [I..n] \) 그리고 \( Q_j, j \in [1..m] \)은 위상, 영상, 운동량 또는 다른 이미지선명자에서 파생한 속성이며 \( a, \beta, \gamma, \delta, \lambda, \mu \)는 오브젝트 특성의 발생도를 표시하는 수치규수이다.

다음 장에서는 특정국적화에 의해 이미지가 분할된 후 이미지데이터베이스에서 연관규칙을 마이닝하는데 효과적인 방법을 찾는다.

4. 다차원 연관규칙 마이닝을 위한 데이터베이스

연관규칙마이닝은 [4,5,6]에서 연구되어 왔다. 루트 X와 여러 Y가 각각 연결된 속성을 시험하는데 X로 설정 규칙에서 만족 \( X \rightarrow Y \) 정상 규칙을 만족 없는 입력의 분리된 속성을 포함하고 있으므로 다차원 연관규칙이 따르다. 관계형 데이터베이스로부터 트플을 포함하는 데이터집합 D에 대해 \( X \rightarrow Y \) 규칙지식은 D에서 트플이 X와 Y양측을 포함한다는 가능성을 가지고 있을 필요가 있다. \( X \rightarrow Y \)의 신뢰도는 트플 T가 X로 포함한다는 조건에 Y로 포함한다는 가능성을 말한다.

\( X \) 속성 집합 \( k \) 연결 속성을 포함하는 집합은 만족 지지도(Support)가 최소지지도 경계보다 크다면 역시 그 값이 크다고 할 수 있다. \( X \rightarrow Y \) 규칙은 최소지지도 최소지지도 경계점 양측을 만족시킨다면 충분히 강하다. 모든 속성이 분리된 속성이름을 가진 규칙은 비반복 속성 다차원연규칙과 밀접하게 연결된다.

데이터마이닝에서는 규칙을 통해 요약된 특성 중요한 특성 또는 메타데이터의 특성을 각각 추출한다. 데이터베이스[7]는 다중 어레이 구조로써 다중차원으로 데이터를 저장하고, 저장된 여러 데이터베이스 정보에서 가장 중요한 주제를 통합하여 보여주는 구조로 이루어져 있는데 이 구조는 이용자가 원하는 트리를 통해 메시지를 찾아오는 역할을 한다. n차원 데이터베이스 \( C[\ldots, A_1, A_n] \)는 \( n \)-D데이터베이스로 \( A_1, A_n \)은 \( n \)차원이다. 큐브의 각 차원에는 \( A_i \) 속성을 표현하고 \( A_i + 1 \) 줄을 포함하며 \( A_i \) 열에서 구별되는 값의 수를 말한다. 정의한 \( A_i \) 줄은 데이터 줄이다. \( A_i \)의 개별값은 각 데이터 줄을 한다. 각 형 줄은 상위층 해당값에 해당하는 카운트의 합계를 저장하는데 사용된다. 큐브의 데이터셀 C \[ a_1, i_1, \ldots, a_n, i_n \]의 초기값 \( r \) (\( A_i = a_1, i_1, A_n = a_n, i_n \), Count) 트플의 해당하는 카운트를 저장한다. 큐브의 정의 셀 C[sum, \( a_2, i_2, \ldots, a_n, i_n \)]에서 * 또는 all 키워드에 의해 표현된 항목은 \( s \) sum을 저장한다. 생성된 퍼즐의 카운트 합계는 \( n \)개성의 \( r \) (\( A_2 = a_2, i_2, A_n = a_n, i_n, Count \))의 두번째에 대해 동일한 값을 공유한다. 개념적으로 데이터베이스는 용량의 경계를 표현한다. \( n \)-D공간(기본육면체)은 모든 데이터 셀로 구성된다. \( (n-1) \)-D 공간은 어떤 차원에서 \( r \) (\( A_i = a_1, i_1, A_n = a_n, i_n, Count \)) 등의 다른 단일 *와 함께 모든 셀을 구성한다. 마지막으로 0-D공간은 \( r(\ast, \ldots, \ast, \text{sum}) \) 값은 \( n \) * 차원으로 한꺼번에 셀을 구성한다. 그림 2는 다차원 데이터베이스의 3D 형식으로 보여준다. 데이터 셀은 데이터를 처리하는데 유연성을 높이면이나 다른 각도에서 데이터를 볼 수 있도록 한다. 합계성은 차원수상에서 정의한 서로 다른 데이터들의 다단계에서 합계할 수 있는 데이터 처리방법을 사용할 수 있도록 하기 때문에 각 차원값에 대한 속성을 갖춘 다차원 데이터베이스를 알리고 구축할 수 있다.[8]

![그림 2. 3D 형태의 데이터모델](image)

4.2 연관규칙 검색 알고리즘

멀티데이터데이터베이스에 데이터베이스 개념을 도입하여 멀티데이터 차원값을 데이터베이스의 속성으로 표현한다. 데이터베이스[7]는 다중 어레이 구조로써 다중차원으로 데이터를 저장하고, 저장된 여러 데이터베이스 정보에서 가장 중요한 주제를 통합하여 보여주는 구조로 이루어져 있는데 이 구조는 이용자가 원하는 트리를 통해 메시지를 찾아오는 역할을 한다. n차원 데이터베이스 \( C[\ldots, A_1, A_n] \)는 \( n \)-D데이터베이스로 \( A_1, A_n \)은 \( n \)차원이다. 큐브의 각 차원에는 \( A_i \) 속성을 표현하고 \( A_i + 1 \) 줄을 포함하며 \( A_i \) 열에서 구별되는 값의 수를 말한다. 정의한 \( A_i \) 줄은 데이터 줄이다. \( A_i \)의 개별값은 각 데이터 줄을 한다. 각 형 줄은 상위층 해당값에 해당하는 카운트의 합계를 저장하는데 사용된다. 큐브의 데이터셀 C \[ a_1, i_1, \ldots, a_n, i_n \]의 초기값 \( r \) (\( A_i = a_1, i_1, A_n = a_n, i_n \), Count) 트플의 해당하는 카운트를 저장한다. 큐브의 정의 셀 C[sum, \( a_2, i_2, \ldots, a_n, i_n \)]에서 * 또는 all 키워드에 의해 표현된 항목은 \( s \) sum을 저장한다. 생성된 퍼즐의 카운트 합계는 \( n \)개성의 \( r \) (\( A_2 = a_2, i_2, A_n = a_n, i_n, Count \))의 두번째에 대해 동일한 값을 공유한다. 개념적으로 데이터베이스는 용량의 경계를 표현한다. \( n \)-D공간(기본육면체)은 모든 데이터 셀로 구성된다. \( (n-1) \)-D 공간은 어떤 차원에서 \( r \) (\( A_i = a_1, i_1, A_n = a_n, i_n, Count \)) 등의 다른 단일 *와 함께 모든 셀을 구성한다. 마지막으로 0-D공간은 \( r(\ast, \ldots, \ast, \text{sum}) \) 값은 \( n \) * 차원으로 한꺼번에 셀을 구성한다. 그림 2는 다차원 데이터베이스의 3D 형식으로 보여준다. 데이터 셀은 데이터를 처리하는데 유연성을 높이면이나 다른 각도에서 데이터를 볼 수 있도록 한다. 합계성은 차원수상에서 정의한 서로 다른 데이터들의 다단계에서 합계할 수 있는 데이터 처리방법을 사용할 수 있도록 하기 때문에 각 차원값에 대한 속성을 갖춘 다차원 데이터베이스를 알리고 구축할 수 있다.[8]
용하는 기호정의는 다음과 같다. \( p \) 는 예시된 속성 수 또는 규칙 \( (p \leq n' \ n \)은 규칙의 차원\)에서 발생하는 속성 변수이다. \( L_k \)는 큰 \( k \) 속성집합의 세트고 \( L_k \)의 각 세트는 속성 집합과 지지 집합, 두가지 필드를 가진다. \( Q \)는 모든 강한 연관규칙의 집합이다. \( \min_{sup} \)는 최소지지도이며 \( \min_{conf} \)는 최소신뢰도이다.

**n-D 규칙검색 알고리즘 (n차원 데이터큐브로부터 다중연관규칙 마이닝)**

**Input:**
1. 규칙 \( R \)은 \( p \) 구분 속성을 포함;
2. 전체산된 차원공간을 가진 \( n-D \) 데이터 큐브, \( C = \{ A_1, \ldots, A_n \} \)
3. \( \min_{sup} \)와 \( \min_{conf} \) 경계점

**Output:** 강한 연관규칙 집합 \( R \)

**Method:**
1. \( p-D \) 셈의 셈 함수를 실행, 만약 셈함계가 \( \min_{sup} \)를 만족시키면 해당하는 \( p \) 속성 집합을 \( L_p \)에 더한다.
2. \( R= \text{rule gen}(L_p, C, \min_{conf}) \);
3. 루턴 \( (R) \);

**근거:** 데이터 큐브의 요약층이 생성되면 \( p-D \) 셈을 실행함으로써 \( L_p \)를 적절 찾아올 수 있다. 만약 전체 데이터큐브가 가능 메모리보다 크다면 \( p-D \) 요약층만이 로드된다. 그다음 \( p \)-속성셋이 주어지면 규칙 생성 방법이 사용된다.

5. 결론
본 논문에서는 멀티미디어데이터를 데이터마이닝하기 위한 프로토타입 시스템을 설계하고 데이터간의 연관규칙을 찾아내기 위해 특성국가행렬을 이용한 이미지분할을 적용하고 특히 데이터베이스로부터 연관규칙을 추출하기 위한 방법으로 다차원 멀티미디아데이터 데이터큐브와 검색 알고리즘을 제안하여 지식을 추출하는 방안을 제안했다. 특히 칼라나 척각 대신 소화(locate)으로 특성을 지정하여 연관을 발생하는 방법을 적용한 멀티미디어 연관규칙은 높은 신뢰도로 데이터를 마이닝하는데 적용될 수 있다. 향후 계속되어야 할 연구로는 이미지 분할을 이용한 내용기반 데이터검색시스템의 세부기능 즉 이미지의 명확한 성분추출 프로세스 기술, 정교한 데이터마이닝을 위한 데이터의 상세화(Granularity) 가능성을 등의 연구를 통해 시스템의 실제 구축 및 데이터큐브 검색 알고리즘의 성능 테스트 그리고 데이터큐브와 데이터마이닝 모델을 개선하는 방법들이 남겨져 있다.

참고문헌


