2001년 한국정보처리학회 추계 학술발표논문집 제8권 제2호

멀티미디어 데이터의 다차원 연관규칙 마이닝

김진옥, 황대준
성균관대학교 전기전자컴퓨터공학부
e-mail : jinny@ece.skku.ac.kr, djhwang@ece.skku.ac.kr

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 마이닝에 대한 관심이 증가하고 있다. 본 논문에서는 특성국지화를 이용한 내용기반의 정보검색 기술과 다차원 데이터큐브 구축기술을 통해 멀티미디어 데이터에서 연관규칙을 찾아내는 멀티미디어 데이터마이닝 시스템 프로토타입을 제안한다. 특히 멀티미디어 데이터의 칼라,질감 등 거시적인 이미지 성분 대신 이미지의 영역성과 유사성을 이용한 특성국지화방법을 이용하여 이미지를 분할함으로써 방대한 데이타에서 효과적인 내용기반의 정보 검색을 시행하고 검색한 벡터를 메타데이타로 한 데이스베이스를 구축한다. 그리고 데이터베이스에서 데이터간 연관규칙을 찾아내어 지식을 마이닝하는데 효과적인 다차원 데이터큐브를 구축하고 여기에 연관규칙 검색 알고리즘을 적용한다.

1. 서론

최근 인터넷과 웹기술의 발전 그리고 이를 기반으로 하는 멀티미디어 컨텐츠가 대량으로 쏟아지고 있다. 특히 웹기술의 발전과 이용자 수 증가에 따라인 멀티미디어 데이터베이스를 추출하여 필요한 지식을 마이닝 하고자 하는 연구가 계속되고 있다[1]. 멀티미디어 데이타마이닝은 멀티미디어 데이타베이스에 저장된 패턴 뿐만 아니라 멀티미디어 데이터관계, 함축적 지식의 추론을 이끌어내는 분야로 데이터마이닝의 하부영역이다. 현재 멀티미디어 데이터를 표현, 저장하고 인덱스하거나 검색하는 기법에 대한 연구는 다양하게 제안되고 있으므로 기존 멀티미디어연구분야와 데이터가 저장된 리파지토리부터 개개인에게 필요한 정보를 지식으로 표현하는 마이닝기술의 결합을 통해 멀티미디어 데이타마이닝 시스템이 구현된다. 본 논문에서는 이미지프로세싱과 데이터 마이닝 기술을 결합한 멀티미디어 기존 데이터마이닝 시스템 프로토타입을 설정하고, 특히 멀티미디어 데이터간의 연관규칙을 찾아내어 지식을 생성할 수 있는 방법을 제안한다. 여기에 적용하는 방법으로 멀티미디어 데이터를 추출하여 데이터베이스화하는 과정에 효율적인 이미지 특성국지화(feature localization)와 연관규칙 마이닝을 위한 다차원 데이터큐브 구축 및 검색알고리즘에 대해 설명한다.

2장에서는 멀티미디어 데이터마이닝 시스템의 프로토타입 디자인을 제안하며 3장에서는 내용 기반 이미지검색 시스템 중 특성국지화 방법에 대해 논하고 4장에서는 연관규칙 마이닝을 위한 데이터큐브 구축방안과 다차원 데이터큐브를 이용한 검색알고리즘에 대해 설명한다. 마지막으로 5장에서는 결론과 향후 연구방향에 대해 설명한다.

2. 멀티미디어 데이터마이닝 시스템 프로토타입

멀티미디어 데이터 마이닝은 컴퓨터비젼, 칼라인식, 이미지프로세싱, 이미지분류, HCI(Human Computer Interface)등 이미지처리분야 기술과 마이닝분야 기술을 총합하여 이루어내는 복합학문영역으로 이미지데이터를 내용기반으로 추출하는 내용기반 이미지 추출시스템과 OLAP(On-Line Analytical Processing)기술, 그리고 데이터간의 특징화, 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 검색 에이전트가 가져온 멀티미디어 데이타를 내용기반의 컨텐츠로 필터링하고 변환하여 압축하는 전처리기모듈과 전처리기를 통해 의미있는 정보만을 추출하여 개별적인 속성을 가진 데이터요약과 정보추출내용으로 구성한 데이터베이스와 데이타베이스의 개별속성을 이용하여 구축한 멀티미디어 데이타큐브에서 드릴다운(Drill-down), 롤업(Roll-up), 슬라이스(Slice), 다이스(Dice)와 같은 OLAP 분석 절차를 거쳐서 데이터간의 상호연관 관계를 찾아내고 분류하며 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.

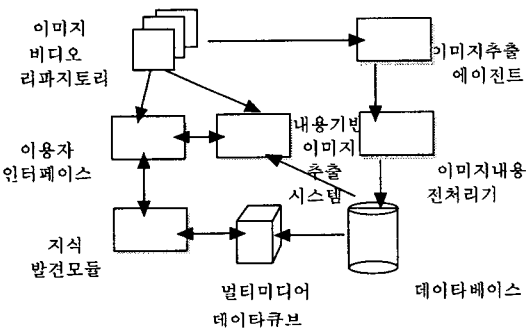

그림 1. 멀티미디어 데이터마이닝 시스템의 구조

제안된 멀티미디어 데이타마이닝 시스템은 웹 자체를 이미지와 비디오 데이터의 저장소로 본다. 이곳에서 대용량 멀티미디어 데이터를 추출하여 처리하는 내용기반 이미지추출 시스템을 이용하여 데이터 검색과 처리 그리고 데이터베이스 구축을 한다. 멀터미디어 데이터처리를 위해 적용되는 내용기반 이미지 추출시스템은 이미지 추출에이전트, 이미지 내용 전처리기, 이용자 인터페이스, 데이터베이스에서 이미지, 비디오특성을 쿼리와 매칭하는 검색커널으로 구성되어 있다.

3. 멀티미디어 데이터의 특성국지화 (Feature localization)

기존 내용기반 이미지 검색기술에 쓰여지는 대부분의 기법은 색이나 질감과 같은 거시적인 이미지 특성에 의존한다. 이 거시적인 방법은 이미지 전체에서 추출된 단순한 통계치에 기반하고 있는데 값을 얻기 쉽고 데이터베이스에 쉽게 저장할 수 있고 또한 값 매칭에는 시간이 거의 걸리지 않는다는 장점이 있다. 그러나 거시적인 방법으로는 특정 대상체의 위치나 대상체의 속성(크기,위치,방향등)을 식별하는 것이 어렵다. 보통 이미지에서 색과 질감 특성을 이용하여 내용검색을 시도하지만 이미지의 영상적 특성국지화(localization), 공간적 상관관계 그리고 비디오일 경우 시간차에 따른 영상의 움직임등의 특성은 간과하고 있다. 따라서 본 논문에서는 전통적인 이미지분할방법이 그 대중성에도 불구하고 이미지와 비디오데이타베이스에서 내용기반의 검색을 위해 적절한 이미지 전처리 방법이 아니라는 것을 표방하면서 그 대신 특징국지화에 대한 새로운 시도를 제안하고자 한다.

특성국지화는 [3]에서 연구한 이미지분할의 새로운 개념으로 이미지 분할은 이미지를 연결되지 않는 지역으로 나누는 과정을 말한다.
이미지분할은 전체이미지를 분리된 영역으로 분할하는 과정이다. 이 영역은 어떤 특성 즉 유사색(회색레벨 세기) 유사질감 등 어떤 속성을 공유하는 픽셀 집합으로 구성된다.
전통적인 분할알고리즘은 다음과 같은 내용을 포함하고 있다. [3]에 의거하여 R 이 지역이라면 (1) 각 영역은 대부분 결합되어 있으며 (2) 각 영역은 분해된다. ($R_i \cap R_j = \emptyset$, $i \neq j$) (3) 분할은 어떤 픽셀이 어떤 영역에 할당되면 완료되고 모든 영역의 총합은 이미지전체이다 ($U_{k=1}^{m} R_k = I$).
이미지분할은 이미지내용을 표현하는데 적절한 방법이 아니다. 더 유용한 방법은 영역성과 유사성으로 특징을 구분하는 특성국지화방법이다
[3]에서 정의한 대로 소역 (locale)은 이미지단계에서 가장 작은 단위로 정방향 픽셀을 이용한다
소역 (locale) ι_x 은 특성 x의 구역이다. ι_x는 ι_x의 지역성을 표현하는 타일집합인 외장 L_x와 mass $M(\iota)_x$ (L_x에서 픽셀 수), centroid $C(\iota)_x$ (L_x에서 픽셀들의 중심), 편차 $\sigma^2(\iota)_x$ (L_x의 픽셀에서 중심까지의 Cartesian 편차), 소역의 모양 등 몇가지 기하학적 계수를 가지고 있다.
타일은 외장을 위한 설정단위이다. 타일은 관련 픽셀수가 충분히 붉으면 '붉다'라고 본다. 타일은 픽셀이 붉거나 약간 푸른색을 띠면 붉을 수도 있고 푸른색일수도 있다. 픽셀이 이미지분할에 대한 설정단위라면 타일은 특성국지화에 대한 설정단위이다. 특성국지화는 중첩이 가능한 거친 분할이라고 할 수 있기 때문에 일치성이 필수적인 방법은 아니다. 특성국지화 방법에서는 이미지 간 위상관계 추정을 위해 소역(locale) 주위의 최소경계 원을 정의하여 이 원과 소역간의 관계를 비교하여 데이터를 검색할 수도 있다. 제안된 특성국지화에서는 근사치 위치가 확인되는 것이지 픽셀이 어떤 영역에 속하느냐를 나타내지는 않는다. 모형매칭과 같은 단순한 프로세스만이 적용된다면 픽셀과 영역간의 관계가 중요하지만. 대량 이미지와 비디오가 처리되는 내용기반의 이미지 검색 영역에서는 단순하고 명확한 매치 자체가 이루어질 가능성이 높지는 않다. 대신, 여러 가지 특성과 그들의 공간적 관계와 관련한 프로세스

가 더 명확하게 이루어진다. 그렇기 때문에 특정 소역(locale)은 추출하기 쉽고 이에 따라 특성국지화 방법이 픽셀에 의해 형성된 상세 영역보다 추출하기 쉽게 된다.

특성국지화를 통해 분할된 이미지는 각 계수값으로 메타데이타형태로 데이터베이스에 저장된다. 그리고 이용자의 쿼리에 따라 연관규칙에 의해 추출된다. 멀티미디어 연관규칙은 데이타베이스에서 특성국지화계수와 같은 영상오브젝트의 특성을 연결하는 규칙을 가지고 있다.

$$\alpha P_1 \wedge \beta P_2 \wedge \ldots \wedge \gamma P_n \rightarrow \delta Q_1 \wedge \lambda Q_2 \wedge \ldots \wedge \mu Q_m \ (C\%)$$

여기서 C%는 규칙의 신뢰정도이며 P_i, , $i \in [1...n]$ 그리고 $Q_j, j \in [1...m]$ 은 위상,영상,운동학상 또는 다른 이미지설명자에서 파생한 속성이고 $\alpha, \beta, \gamma, \delta, \lambda$ 는 오브젝트 특성의 발생정도를 표시하는 정수계수라고 본다.

다음 장에서는 특성국지화에 의해 이미지가 분할된 후 이미지데이타베이스에서 연관규칙을 마이닝하는데 효과적인 방법을 찾는다.

4. 다차원 연관규칙 마이닝을 위한 데이타큐브

연관규칙마이닝은 [4,5,.6]에서 연구되어 왔다. 몸체 X와 머리 Y가 각각 연결된 속성세트를 구성하는 X⇒Y 형성 규칙에서 만약 {X, Y}가 한개 이상의 분리된 속성을 포함하고 있으면 다차원 연관규칙이라고 본다. 관계형 데이터베이스로부터 튜플을 포함하는 데이터집합 D에 대해 D에서 X⇒Y 규칙지원은 D에서 튜플이 X와 Y양측을 포함한다는 가능성을 가지고 있음을 내포한다. X⇒Y의 신뢰도는 튜플이 X를 포함한다는 조건하에 Y를 내포한다는 가능성을 말한다.

K 속성 집합(k 연결 속성을 포함하는 집합)은 만약 지지도(Support)가 최소지지도 경계보다 크다면 역시 그 값도 크다고 할 수 있다. X⇒Y규칙은 최소지지도과 최소신뢰도 경계점 양측을 만족시킨다면 충분히 강하다. 모든 속성이 분리된 속성이름을 가진 규칙을 비반복 속성 다차원연관 규칙이라고 한다.

데이터마이닝에서는 쿼리를 통해 요청된 태스크 관련 데이터를 추출하는데 최소지지도 경계는 여기서 너무 많은 개별값을 가진 속성을 제거하여 적절한 차원을 추출하는데 이용되거나 낮은 데이터값을 필요한 값으로 정제하는데 사용된다.

4.1 다차원 데이터큐브

멀티미디어 데이타베이스에 데이터큐브 개념을 도입하여 멀티미디어 차원값을 데이터큐브의 속성으로 표현한다. 데이터큐브[7]는 다중 어레이 구조로써 다중차원으로 데이터를 저장하고, 저장된 여러 데이터 레벨 정보에서 가장 중요한 주제를 통합하여 보여주는 구조로 이루어지는데 이 구조는 이용자가 던지는 쿼리를 통해 매치값을 찾아주는 역할을 한다. n차원 데이타큐브 C[... A_1 A_n]는 n-D데이터베이스로 A_1, A_n은 n차원이다. 큐브의 각 차원에서는 A_i 는 속성을 표현하고| A_i|+1 줄을 포함하는데| A_i|는 A_i 차원에서 구별되는 값의 수를 말한다. 첫번째| A_i| 줄은 데이터 줄이다. A_i의 개별값은 한 개 데이터 줄을 취한다. 합계 줄인 마지막 줄은 상위줄 해당칼럼에 해당하는 카운트의 합계를 저장하는데 사용된다. 큐브에서의 데이터셀인 C [a_1, i_1,∴, a_n, i_n] 는 초기관계 r (A_i = a_1, i_1, A_n = a_n, i_n, *Count*) 튜플에 해당하는 카운트를 저장한다. 큐브의 합계 셀 C[*sum*, a_2,, i_2,, ... , a_n , i_n]에서 * 또는 all 키워드에 의해 표현되는 합계는 sum_1을 저장한다. 생성된 튜플의 카운트 합계는 n번째 칸 r (A_2 = a_2, i_2, A_n = a_n, i_n,, *Count*)의 두번째에 대해 동일한 값을 공유한다. 개념적으로 데이터큐브는 육면체의 격자로 표현된다. 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*) 등과 같은 단일 * 와 함께 모든 셀을 구성한다. 마지막으로 *0-D* 공간은 r(*,*,...,*, sum_n) 과 같은 n * d 차원으로 한가지 셀을 구성한다. 그림 2는 다차원 데이터큐브를 3D 형태로 보여준다, 데이터 큐브는 데이터를 처리하는데 유연성을 줄 뿐만 아니라 다른 각도에서 데이터를 볼 수 있도록 한다. 합계셀은 차원속성에서 정의한 서로 다른 레벨의 다단계에서 합계값을 빨리 처리해낼 수 있도록 하기 때문에 각 차원도메인에 모든 속성값을 합친 다차원 데이터큐브를 빨리 구축할 수 있다[8].

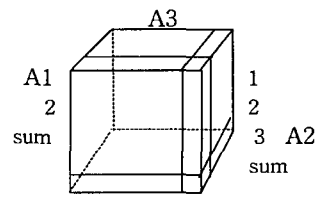

그림 2. 3D 형태의 데이타모델

4.2 연관규칙 검색 알고리즘

멀티미디어의 다차원 연관규칙에 대한 큐브기반 검색 알고리즘은 전계산된 데이터큐브에서 마이닝이 이루어진다는 가정하에 실행된다. 알고리즘에서 사

용하는 기호정의는 다음과 같다. p는 예시된 속성 수 또는 규칙 ($p \leq n$(n은 큐브의 차원))에서 발생하는 속성 변수이다. L_k 는 큰 k 속성집합의 셋이고 L_k의 각 멤버는 속성 집합과 지지 집합, 두가지 필드를 가진다. R은 모든 강한 연관규칙의 집합이다. *min_sup* 는 최소지지도이며 *min_conf* 는 최소신뢰도이다.

n-D 큐브검색 알고리즘(n차원 데이타큐브로부터 다중연관규칙 마이닝)

Input : (1) 규칙 R은 p 구분 속성을 포함;
(2) 전계산된 차원공간을 가진 n-D 데이터 큐브, C [A_1 , ..., A_n]
(3) *min_sup*와 *min_conf* 경계점

Output : 강한 연관규칙 집합 R

Method : 1. p-D셀의 셀 합계를 실행. 만약 셀합계가 *min_sup*를 만족시키면 해당하는 p-속성 집합을 L_p에 더한다
2. R= *rule_gen*(L_p, C, *min_conf*);
3. Ruturn { R };

근거 : 데이터 큐브의 요약층이 계산되면 p-D셀을 실행함으로써 L_p를 직접 찾을 수 있다. 만약 전체 데이터큐브가 가능 메모리보다 크다면 p-D 요약층만이 로드된다. 그다음 p-속성셋이 주어지면 규칙 생성 방법이 사용된다.

5. 결론

본 논문에서는 멀티미디어데이타를 데이터마이닝하기 위한 프로토타입 시스템을 설계하고 데이터간의 연관규칙을 찾아내기 위해 특성국지화방법을 이용한 이미지분할을 적용하고 특히 데이터베이스로부터 연관규칙을 추출하기 위한 방법으로 다차원 멀티미디어데이타 데이터큐브와 검색알고리즘을 제안하여 지식을 추출하는 방안을 제안했다. 특히 칼라나 질감 대신 소역(locale)으로 특성을 지역화하여 내용을 분할하는 방법을 적용한 멀티미디어 연관규칙은 높은 신뢰도로 데이터를 마이닝하는데 적용될 수 있다. 향후 계속되어야 할 연구로는 이미지 분할을 이용한 내용기반 데이타검색시스템의 세부기능 즉 이미지의 명확한 성분추출 프로세싱 기술, 정교한 데이터마이닝을 위한 데이터의 상세화(Granularity) 가능성 등의 연구를 통한 시스템의 실제 구축 및 데이터큐브 검색 알고리즘의 성능 테스트 그리고 데이터큐브와 데이터마이닝 모듈을 개선하는 방법등이 남겨져 있다.

참고문헌

[1] I. Bhandari, E. Colet, J. Parker,Z,Pines, R. Pratap, Advanced scout: Data Mining and knowledge discovery in NBA data. Data Mining and Knowledge Discovery,(191): 121-125, 1997.

[2] S. Chaudhuri, U. Dayal. An overview of data warehousing and OLAP technology.SIGMOD Record 26:65-74, 1997.

[3] Z. N. Li, O. R. Zaiane, and Z. Tauber, Illumination Invariance and object model in content -based image and video retrieval. Journal of Visual Communication and Image Representation,10(3): 219-244, Sept., 1999.

[4] M. J. Egenhofer and J. Sharma, Topolical relations between regions in r r^2 and z^2. In Advances in Spatial Databases(SSD'93), Singapore, 1993.

[5] R. Ng, L. V. S. Lakshmana, J. Han, and A. Pang, Exploratry mining and pruning optimizations of constrained associations rules. In Proc. ACM-SIGMOD, seattle, 1998.

[6] R. Srikant and R. AGrawal, Mining quantitative assocation rules in large relational tables. In Proc. ACM-SIGMOD,pages1-12, Montreal, 1996

[7 R. Kimber, J. Han, J. Y. Chiang, *Metarule-guided mining of multi-dimensional* association rules using data cubes, In In Proc.3rd Int. Conf. Knowledge Discovery and Data Mining(KDD'97), pages 207-210, California, August, 1997.

[8] Y. Zhao, P. M. Deshpande, J. F. Naughton. An array-based algorithm for simultaneous multidimensional aggregates. In Proc. ACM-SIGMOD Int. Conf. Management of Data, pages 159-170. 1997.

[9] V. Athitsos, M. J. Swain and C. Frankel, Distinguishing photographs and graphics on the world wide web, In Proc. IEEE Workshop on Content-Based Acess of Image and Video Libraries,1997.

[10] Y. Rubner, W. Chang, A. Zhang, Semantic Clustering and querying on heterogeneous features for visual data, In Proc. ACM Multimedia, pp.3-12, Bristol, UK, 1998.

[11] J. Z. Wang, J. Li, R. M. Gray, G. Wiederhold, Unsuperivsed multiresolution segementation for images with low depth of field, IEEE Trans.on PAMI vol. 23, no.1, pp 85-91, 2000.