제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

빈발 항목과 의미있는 희소 항목을 포함한 이미지 데이터 연관 규칙 마이닝

송임영, 석상기
서울산업대학교 컴퓨터공학과
e-mail: {siy0614, sksuk}@snut.ac.kr

Association Rules Mining on Image Data with Recurrent Items and Significant Rare Items

Im-Young Song, Sang-Kee Suk
Dept. of Computer Science and Engineering, Seoul National University of Technology

요 약

최근 인터넷과 웹 기술의 발전 그리고 이를 기반으로 하는 다양한 멀티미디어 컨텐츠가 홍수를 이루고 있지만 멀티미디어 데이터에서 체계적으로 연관 규칙을 마이닝 하는 연구는 초기 단계이다. 본 논문에서는 이미지 프로세싱 분야 및 내용 기반 이미지 검색에 대한 기존 연구를 바탕으로 이미지 데이터 저장소에 저장된 재생성 항목과 희소하게 발생하지만 상대적으로 특정 항목과 높은 비율로 동시에 나타나는 희소 항목을 포함한 내용기반의 이미지 연관 규칙을 찾아내기 위한 탐사 기법을 제안한다. 실험 결과 제안된 알고리즘은 기존의 재생성 항목만을 고려한 알고리즘보다 희소 항목을 포함하여 연관 규칙을 탐사하므로 같은 종류의 이미지가 모여 있는 저장소에서 이미지 오브젝트간의 연관 관계를 발견하는 이미지 데이터 마이닝에 효과적이다.

1. 서론

웹 기술의 발전과 이용자 수 증가에 따라 멀티미디어 데이터베이스를 추출하여 필요한 지식을 마이닝하고자 하는 연구가 계속되고 있다. 멀티미디어 정보는 문자나 숫자뿐만 아니라 텍스트, 이미지, 오디오, 비디오와 같은 모노 미디어의 내용에 의한 논리적 관계와 시공간 관계, 사용자 상호작용 등에 따라 구성된 정보 형태이다. 멀티미디어 정보를 구성하고 있는 매체 중에서도 이미지 정보는 사람의 시각으로 인지되는 매체로서 멀티미디어 정보 가운데서도 많은 비중을 차지하며 다른 매체에 비해 정보의 전달 효과가 크기 때문에 이미지 매체를 이용한 다양한 응용들이 개발되고 있다.

최근, 이러한 이미지 데이터를 신속하고 효율적으로 검색하기 위하여 다양한 내용 기반 이미지 검색 기법(content-based image retrieval)[1,2,3]들이 제시되었다.

본 논문에서는 전체 이미지 데이터에서는 희소하게 나타나지만 특정 항목과 상대적으로 높은 비율로 동시에 나타나는 항목에 대한 연관성과 이미지 데이터에서 반복적으로 발생하는 재생성 항목에 대한 내용 기반 연관성을 고려한 연관 규칙 탐사 기법을 제안한다.

2. 관련 연구

연관 규칙 마이닝은 최근 데이터 마이닝 분야에서 광범위하게 연구되어 왔으며[4,5,6], 몇몇 제안 알고리즘은 프로세스 할 수 있는 형태로 변환한 영상 데이터 분야에 적용될 수 있지만 영상 데이터가 가진 이미지 정보의 특이성을 찾아내기는 어렵다. 특정 영상 항목은 한 이미지에서 빈번하게 발생할 수 있으므로 기존에 제안된 연관 규칙의 적용은 이미지와 비디오 데이터로부터 연관 규칙을 마이닝하는데 한계가 있다.

이미지에서 반복적으로 발생하는 항목을 고려하여 탐색한 연관 규칙을 재생성 항목의 연관 규칙이라 한다. 동일한 오브젝트들은 이미지에서 반복적으로 발생함을 고려하여 지지도는 이미지의 수보다는 오브젝트의 수로 반영하는 오브젝트 기반 지지도를 사용하였다.

3. 최대 발생 빈도와 의미있는 희소 항목을 포함한 연관 규칙 탐사 기법

기존의 MaxOccur 알고리즘[7]은 이미지 트랜잭션에서 반복적으로 발생하는 항목에 대한 연관 규칙을 탐사해 올 수 있다. 하지만 이미지에서 희소하게 발생하는 항목에 대한 고려가 없다. 본 논문에서는 데이터의 상대 지지도를 이용하여 희소 데이터를 포함한 연관 규칙을 탐사할 수 있는 RSA 알고리즘[8]을 이미지 오브젝트 기반의 지지도로 모델링하고 MaxOccur 알고리즘의 재생성 항목을 고려한 MORSA(Max Occurrence & Relative Support Apriori) 방법을 제안하고 그 알고리즘을 구현하였다.

의미있는 희소 항목과 최대 발생 빈도는 다음과 같이 정의한다.

[정의 1] 의미있는 희소 항목

의미있는 희소 항목이란 전체 이미지 데이터에서 발

생 빈도수가 1차 지지도는 만족하지 못하지만 2차 지지도를 만족하는 희소 항목 중에서 특정 항목들과 높은 비율로 동시에 나타나는 항목이다.

[정의 2] 최대 발생 빈도

1-빈발 항목 집합의 이미지에서 최대 발생 빈도수

3.1 MORSA

기존의 Apriori, MaxOccur 알고리즘들은 빈발 항목 구성 단계에서 하나의 지지도를 이용하여 연관 규칙의 생성에 쓰일 항목들을 전지하고, 지지도를 만족하는 빈발 항목 집합에서 규칙을 생성한 후, 신뢰도를 적용하여 규칙의 타당함을 검증하는 방식으로 진행된다.

그러나, MORSA 방법에서는 2개의 지지도가 설정되는데 값을 만족하는 빈발 항목과 만족하지 못하는 희소 항목으로 구분된다.

[정의 3] 1차 지지도

빈발 항목 탐사 과정에 사용하기 위하여 사용자가 정의한 지지도의 임계값

[정의 4] 2차 지지도

희소 항목 탐사 과정에 사용하기 위하여 사용자가 정의한 지지도의 임계값

중복된 규칙이 생성되거나 희소 항목을 탐사하기 위해서 1차 지지도와 2차 지지도의 설정은 1차 지지도 > *2차 지지도를 만족하도록 설정해야 한다.*

또한, 항목 사이의 상대적인 빈도를 고려하여 연관 규칙을 탐사할 수 있는 상대 지지도를 사용한다. 상대 지지도는 2차 지지도를 만족하는 희소 항목과 임의의 다른 항목 사이의 상대적인 지지도이며, 상대 지지도를 이용해 의미있는 희소 항목들을 탐사할 수 있다.

[정의 5] 상대 지지도

이미지 데이터가 항목 집합 $I=\{i_1, i_2, \dots, i_m\}$과 같이 구성되고 *이미지 항목 i의 지지도가 supp(i)로 표현된다면*, 속성 사이의 상대적인 지지도를 의미하는 준 빈발 항목 후보 집합에서의 상대 지지도 $Rsup(i_1, i_2, \dots, i_k)$의 정의는 아래와 같다.

$$Rsup(i_1, i_2, \dots, i_k) = max(supp(i_1, i_2, \dots, i_k)/supp(i_1), supp(i_1, i_2, \dots, i_k)/supp(i_2), \dots, supp(i_1, i_2, \dots, i_k)/supp(i_k))$$

상대 지지도는 $0 \le Rsup \le 1$인 값으로서 이미지 항목을 구성하는 각각의 항목들이 후보 항목과 이루는 상대 지지도들을 비교하여 이 중 가장 큰 값을 선택한다. 이는 후보 항목을 구성하는 이미지 항목 $i_1, i_2, \dots, i_k$ 각각이 후보 항목 집합 $\{i_1, i_2, \dots, i_k\}$에 대하여 상대적으로 얼마만큼의 비율의 지지도로 나타나는지에 대한 척도이며 사용자는 탐사할 항목의 상대 지지도의 임계값인 최소 상대 지지도(Minimum Relative Support)를 지정하여 규칙을 탐사한다.

최소 상대 지지도와 준 빈발 항목 집합의 정의는 다음과 같다.

[정의 6] 최소 상대 지지도

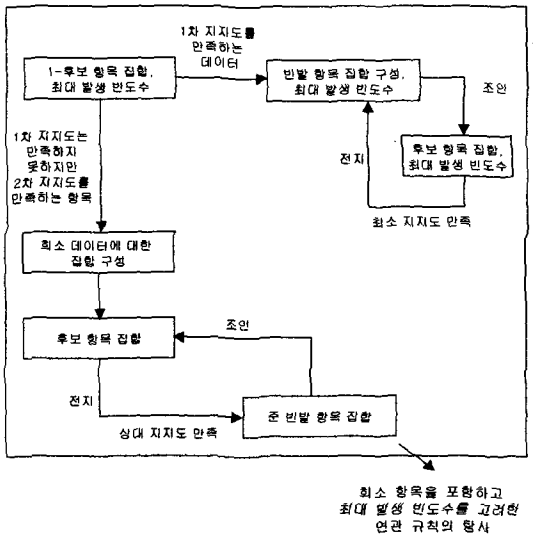
사용자에 의해 정해진 상대 지지도의 임계값으로 *minRsup*로 표기한다.

[정의 7] 준 빈발 항목 집합

준 빈발 항목 집합은 2차 지지도를 만족하는 항목 집합에서 최소 상대 지지도를 만족하는 항목의 집합이다.

상대 지지도는 1차 지지도는 비록 만족하지 못하지만 2차 지지도를 만족하는 희소한 항목을 대상으로 적용된다. 준 빈발 항목 탐사 단계에서 상대 지지도를 계산하며, 계산된 상대 지지도의 값이 최소 상대 지지도를 만족하면 준 빈발 항목이 된다. 최소 상대 지지도 값이 높을수록 사용자는 동시에 나타나는 비율이 큰 항목을 선택함을 의미한다.

MORSA에서의 상대 지지도와 최대 발생 항목 빈도수 사용 과정은 [그림 1]과 같다. 각 과정은 더 이상의 후보 항목을 생성할 수 없을 때까지 반복된다.



[그림 1] 상대 지지도와 최대 발생 빈도수 사용 과정

후보 항목 생성은 희소 항목을 포함한 후보 항목을 구성할 수 있어야 한다.

1차 지지도를 만족하는 원소들은 기존의 MaxOccur 방식과 동일하다. 후보 항목의 구성은 기존 RSA 알고리즘에서와 같이 데이터베이스의 모든 항목에 대해 지지도를 카운트하여 1차 지지도를 만족하는 빈발 항목에 대한 1-후보 항목 집합 C_1과 1차 지지도를 만족하지 못하지만 2차 지지도를 만족하는 희소 항목 집합의 후보 항목 집합 NC_1으로 각각 분리되어 생성된다. 단, morsa_gen에서는 빈발 항목과 희소 항목 집합을 모두 고려한다.

MORSA 알고리즘은 2차 지지도와 최소 상대 지지도라는 척도를 이용하여 비록 전체 이미지에서는 희소하게 나타나지만 특정 항목들과 상대적으로 높은 비율로 발생하는 항목들

과 이미지 데이터에서 반복적으로 발생하는 항목들을 추출해 낼 수 있도록 설계되었다. 알고리즘에 F_k는 MaxOccur 방식과 동일하고 NF_k와 NFF_k는 2-후보 항목의 생성만 다를 뿐 나머지 과정은 동일하므로 NFF_k의 탐사 과정은 생략하였다.

MORSA의 알고리즘은 [그림 2]와 같으며, 사용되는 변수와 함수는 [표 1]과 같다.

[표 1] MORSA에서 사용되는 변수들과 함수들

F_k	빈발 항목
$Support_1$	1차 지지도
$Support_2$	2차 지지도
$minRsup$	최소 상대 지지도
$Rsup$	이미지 항목의 상대 지지도
NC_k NFC_k	회소 항목에 대한 k-후보 항목 집합
NF_k NFF_k	k-준 빈발 항목 집합

```
D : Database
C1 ← {1-후보 항목 집합과 그 지지도}
F1 ← {1-빈발 항목 집합과 그 지지도}
M ← {1-빈발 항목 집합의 이미지에서 최대 발생 빈도}
NF1←{1차 지지도를 만족하지 못하지만 2차 지지도를
      만족하는 희소 항목}

for each item ∈ F1
Fk ← MaxOccur

if(k=2) {
        NC2 ← NF1 ⋈ NF1
        NFC2 ← NF1 ⋈ F1
}
for(k ← 3; NFk-1 ≠ ∅, NFFk-1 ≠ ∅; k++) {
    NCk ← morsa_gen(NFk-1, NFk-1)

    for all image  t ∈ D {
        Nct = subset(NCk , t)

        for all candidates nc ∈ NCt
            nc.support ← object_based_support(nc)
        if nc.support ≥ Support2 {
            for each item ik in nc  {
                nc.Rsup(i1, i2, ... , ik) =
                    max(supp(i1, i2, ... , ik)/supp(i1),
                        supp(i1, i2, ... , ik)/supp(i2), ... ,
                        supp(i1, i2, ... , ik)/supp(ik))
            }
          if nc.Rsup ≥ minRsup
            NFk={nc∈NCt|nc.Rsup≥ minRsup}
        }
    }
}
Result ←MaxOccur의 결과 + UkNFk + UkNFFk

object_based_support (nc)
{
    Dk에서 nc 개수 /
          Σ all transaction T에서의 t집합에서
                    k항목의 부분 집합 수
}
morsa_gen(P, Q)
{
    select p.Item1, q.Item1, ... , q.Itemk-1
    from P p, Q q
    where p.Item1 <= q.Item1, .... ,
          p.Itemk-1 <= q.Itemk-1
}
```

[그림 2] MORSA 알고리즘

4. 실험 결과 및 분석

이 장에서는 MORSA를 기존의 연관 규칙 탐사 방법인 Apriori, MaxOccur와 RSA 알고리즘들과 비교하여 평가한다.

제안된 알고리즘의 확장성, 효율성, 성능 등을 평가하기 위하여 15개 정도의 랜덤한 속성을 가진 오브젝트를 포함한 이미지 집합을 생성하였다.

Apriori 알고리즘, RSAA는 트랜잭션 수 기반의 지지도를 이용하기 때문에 본 논문에서 제안한 알고리즘과 비교하기 위해 이미지 트랜잭션 기반의 지지도를 가진 것으로 구현했고, MaxOccur 알고리즘과 MORSA 알고리즘은 이미지 오브젝트 기반의 지지도를 이용했다.

4.1 각 알고리즘별 성능 평가

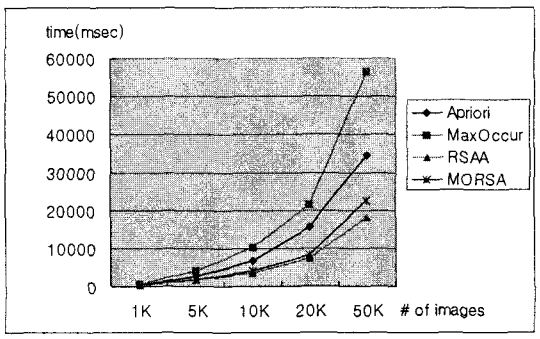
테스트 환경은 Pentium Ⅲ 850, 256MB, Windows 2000, J2SDK 1.4.0 환경에서 수행하였고 랜덤하게 생성된 50,000건의 이미지를 사용하여 각 알고리즘의 성능 평가를 수행하였다.

각 시행 결과는 2회 이상 반복 수행한 후의 결과이다.

[표 2] 이미지 수에 따른 알고리즘별 평균 실행 시간

# of images	Apriori	MaxOccur	RSAA	MORSA
1K	380	621	381	350
5K	2,494	4,125	1,802	1,802
10K	6,659	10,284	3,605	4,196
20K	15,372	21,451	7,350	8,281
50K	34,679	56,451	18,197	22,402

[표 2]는 네 가지 알고리즘에 대한 평균 실행 시간을 보여 준다. Apriori와 RSA 알고리즘은 이미지 트랜잭션 기반의 지지도를 적용하여 Apriori 알고리즘의 최소 지지도는 0.05로, RSA 알고리즘의 1차 지지도는 0.05로 설정하고 MaxOccur 알고리즘과 MORSA 알고리즘은 오브젝트 기반의 지지도를 적용하여 MaxOccur 알고리즘의 최소 지지도는 0.015로, MORSA 알고리즘의 1차 지지도는 0.015로 설정했다. [그림 3]에서와 같이 MORSA 알고리즘의 성능은 1초당 평균 2,000개 이상의 이미지를 처리한다.



[그림 3] 수행 시간 비교 그래프

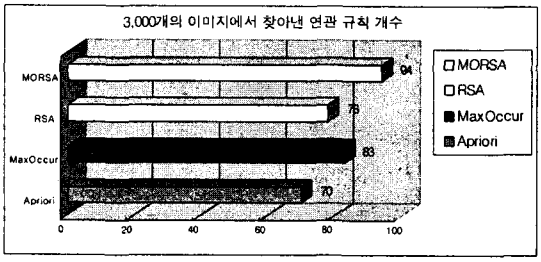
MaxOccur 알고리즘과 MORSA 알고리즘의 차이점은 희소 항목과 재생성 항목을 포함하여 연관 규칙을 탐색한 것이다. 따라서 두개의 지지도로 빈발 항목과 희소

항목을 구분한 후 희소 항목에 대해서는 최소 상대 지지도를 만족하는 항목들을 추출해 낼 수 있으므로 불필요한 규칙을 제거할 수 있고 성능면에서 더 우수함을 보인다.

[표 3] MORSA 알고리즘의 지지도별 수행 시간
(단위 msec)

# of images	0.025	0.02	0.015	0.01
1K	260	281	331	511
5K	1,322	1,402	2,133	2,854
10K	2,714	2,744	4,176	6,129
20K	5,338	5,698	8,573	12,017
50K	13,910	14,882	22,102	32,547

[표 3]은 다른 지지도별 MORSA 알고리즘의 평균 수행 시간으로 지지도가 낮아질수록 수행 시간이 더 소요됨을 알 수 있다.



[그림 4] 알고리즘별 빈발 항목 집합 검색 결과

[그림 4]는 3,000개의 이미지에서 각 알고리즘별 찾아낸 연관 규칙 개수이다. 본 논문에서 제시한 MORSA 알고리즘은 MaxOccur알고리즘과 같은 오브젝트 기반의 지지도를 사용하지만 두 개의 지지도와 최소 상대 지지도를 사용하여 성능 향상과 보다 많은 빈발 항목 집합을 탐색해 내는 것을 확인할 수 있다.

5. 결론 및 향후 연구 과제

본 논문에서는 기존에 발표된 멀티미디어 데이터 마이닝 연관 규칙 알고리즘에서 고려한 재생성 항목과 희소하게 발생하지만 특정 항목과 상대적으로 빈발하게 발생하는 항목을 고려하여 내용 기반의 이미지 연관 규칙을 탐색해내기 위한 MORSA 알고리즘을 제안했다.

제안한 알고리즘은 색상, 모양, 크기, 질감 등의 속성 벡터 값을 이용하여 이미지에서 재생성 되는 항목을 찾고 또한 희소하게 발생하는 항목에 대한 연관 관계를 규명한다.

실험 결과 제안된 알고리즘은 기존의 이미지 연관 규칙 탐사 기법인 MaxOccur 알고리즘 보다 연관 규칙을 탐색하는데 있어서 동일 시간 내에 더 많은 양의 이미지 트랜잭션을 처리할 수 있기 때문에 같은 종류의 이미지가 모여 있는 저장소에서 연관 관계를 발견하는 이미지 데이터 마이닝에 효과적이다.

향후 연구 방향으로는 이미지의 다양한 속성을 표현하는 특징 데이터를 효율적으로 추출하기 위해 알고리즘의 속도 개선과 후보 항목의 전지 기법에 대한 연구가 필요하다.

참고문헌

[1] V. N. Gudivada and V. V. Raghavan, "Content-Based Image Retrieval Systems," IEEE Computer, 28(9), 1995

[2] 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

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

[4] R. Agrawal and R.Srikant, "Fast Algorithms for Association Rules," in Proceedings of the 20th VLDB Conference, Santiago, Chile, Sept., 1994

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

[6] 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.

[7] O. R. Zaïane, J. Han, and H. Zhu, "Mining Recurrent Items in Multimedia with Progressive Resolution Refinement", Proc. 2000 Int. Conf. on Data Engineering (ICDE'00), San Diego, CA, March 2000.

[8] 하단심, 황부현, "상대 지지도를 이용한 의미 있는 희소 항목에 대한 연관 규칙 탐사 기법", 정보과학회지 제 28권, 2001