# 공간 분할 지수를 이용한 이미지 데이터 연관 규칙 마이닝

송임영° 김경창 석상기
홍익대학교, 서울산업대학교
{iysong°, kckim}@cs.hongik.ac.kr, sksuk@snut.ac.kr

# Association Rules Mining of Image Data using Spatial Factor

ImYoung Song° Kim K.C. Suk S.K.
Hongik University, Seoul national University of Technology

요 약

본 논문에서는 기존의 멀티미디어 연관 규칙 알고리즘인 Max occur 알고리즘에서 추출한 빈발 항목 집합의 결과들에 대하여 빈발 항목 집합들끼리의 공간적인 연관 관계를 고려하기 위해 공간 데이터 마이닝의 대표적인 공간 분할 방법인 그리드 셀 기반으로 공간 분할 지수(spatial facotr)인 SF를 이용한 이미지 공간 연관 규칙 마이닝 방법을 제시한다. 또한 최소 공간 지지도를 적용하여 이미지 데이터에서 반복적으로 발생하는 항목과 항목간의 공간 관계를 통해 이미지 연관 규칙을 마이닝 하는데 보다 유효한 알고리즘을 제안한다.

## 1. 서 론

공간 데이터 마이닝은 기존의 데이터 마이닝에 공간 개념을 추가하여 확장한 것으로, 공간 데이터에 대한 지식 탐사 과정으로 정의할 수 있다. 그리고 공간 데이터 마이닝을 통해 공간 패턴 간 객체간의 연관 관계 등을 얻을 수 있다[1]. 공간 연관 규칙은 공간 데이터베이스 안에 존재하는 하나 또는 그 이상의 공간 객체들과 또 다른 공간 객체들 간의 밀접한 관계를 서술하는 규칙이다.

본 논문에서는 이미지 데이터에서 반복적으로 발생하는 항목과 항목간의 공간 관계를 통해 이미지 연관 규칙을 마이닝 하는데 보다 유효한 알고리즘을 제안한다.

본 논문의 구성은 2장에서는 연관 규칙과 공간 데이터 마이닝에 대하여 기술하고, 3장에서는 본 논문에서 제안하는 공간 관계를 갖는 이미지 연관 규칙 탐사 방법을 기술하고, 4장에는 실험 결과로 기존의 알고리즘과의 성능을 비교하고 분석하였으며 5장은 결론 및 향후 연구를 제시한다.

## 2. 관련 연구

### 2.1 이미지 빈발 항목의 연관 규칙

연관 규칙 마이닝은 최근 데이터 마이닝 분야에서 광범위하게 연구되어 왔으며, 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.

### 2.2 공간 데이터 마이닝

공간 데이터베이스는 공간 데이터와 비공간 데이터로 구성이 되어 있다. 공간 데이터는 특정한 공간에 위치하는 객체와 관련된 데이터로써 거리, 위상에 대한 정보를 담고 있다[1,2]. 비공간 데이터는 공간 데이터를 설명해 주는 데이터이다. 실세계의 공간 객체는 "위치"와 같은 공간 데이터와 "이름", "전화번호"와 같은 비공간 데이터가 결합하여 표현된다.

### 2.3 공간 연관 규칙 탐사

공간 연관 규칙은 기존의 연관 규칙을 공간 데이터 마이닝에서도 사용할 수 있도록 확장한 것으로 공간 데이터베이스에서 다른 집합의 특징을 통해 하나 혹은 어떤 집합의 특징에 대한 관련성 혹은 관계를 기술하는 것이다[3]. 공간 연관 규칙은 X→Y로 표현되며, X, Y는 adjacent_to, near_by, inside, close_to, intersecting과 같은 공간 관계를 설명하는 공간 관계 술어가 포함된 조건식의 집합으로 정의된다. 예를 들어 「is_a(x, gas_station) → close_to(x, highway) (75%)」라는 규칙은 gas_station이 highway와 근접할 확률이 75%라는 공간 연관 규칙이다. 공간 연관 규칙은 기존의 연관 규칙 탐사 방법과 유사하게 지지도와 신뢰도를 이용하여 탐사되며 탐사 결과로 공간 데이터와 공간 데이터 간의 공간 관계에 대한 공간 연관 규칙을 얻을 수 있다.

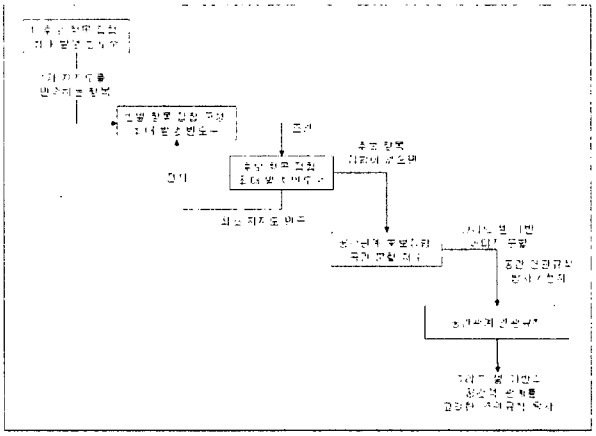
## 3. 공간 분할 지수를 이용한 이미지 데이터 연관 규칙 마이닝

본 논문의 목적은 이미지에 포함된 오브젝트간의 공간 관계를 이용해 데이터간의 보다 정확한 공간 연관 규칙을 찾아내는 것이므로 이 장에서는 멀티미디어 연관 규칙을 정의한 MaxOccur 알고리즘[4]으로 발견된 항목 집합들에 대하여 이미지를 일정한 크기의 셀로 나누어 이미지 데이터에서 반복적으로 발생하는 빈발 항목들에 대한 셀들 간의 위치적인 관계를 이용하여 내용 기반 연

관성을 고려한 공간 연관 규칙 탐사 기법을 제안한다.

### 3.1 GISAR 연관 규칙 탐사 과정

GISAR(Grid cell based on Image Spatial Association Rules) 알고리즘을 사용하여 그리드 셀 기반 공간 연관 규칙 탐사 과정을 보면 (그림 1)과 같다.



(그림 1) 그리드 셀 기반 공간 연관 규칙 탐사 과정

기존 멀티미디어 연관 규칙 알고리즘을 적용하여 나온 빈발 항목 집합들이 본 논문에서 제안하는 알고리즘의 공간 관계 후보 항목 집합들이 된다. 공간 관계 후보 항목 집합들에 *SF*로 이미지를 그리드 셀로 분할하여 공간 연관 규칙을 탐사, 전지하여 공간 관계 연관 규칙을 구한다.

### 3.2 GISAR 알고리즘

GISAR 알고리즘은 Max Occur를 통하여 추출된 빈발 항목 집합을 후보 항목으로 하여 그리드 셀 구조를 기반으로 한 셀 관계 연산을 통하여 객체들 간의 거리 계산 대신에 셀들 간의 관계를 이용하여 공간 관계를 따져 이미지 연관 규칙을 마이닝 하는데 보다 유용한 알고리즘으로 설계되었다. GISAR의 알고리즘은 (그림 2)와 같다. 아래 알고리즘에서 사용되는 자료 구조와 함수의 역할은 <표 1>과 같다.

<표 1> GISAR에서 사용되는 자료 구조와 함수들

| | |
|---|---|
| $F_K$ | 빈발 항목 |
| $support_1$ | 1차지지도 |
| $support_2$ | 2차지지도 |
| *SFC* | 공간관계 후보 집합 |
| *SF* | 공간지수 |
| *SFS* | 공간관계 빈발 항목집합 |
| *MSsup* | 최소공간지지도 |
| *Ssup* | 공간지지도 |

### 4. 실험 결과 및 분석

이 장에서는 빈발 항목 집합들끼리의 공간적인 연관 관계를 고려하여 본 논문에서 제안하고자하는 빈발 항목 집합들의 공간적인 관계를 살펴 봄 으로써 이미지에서 빈발한 오브젝트들 간의 공간적인 관계를 살펴 볼 수 있다.

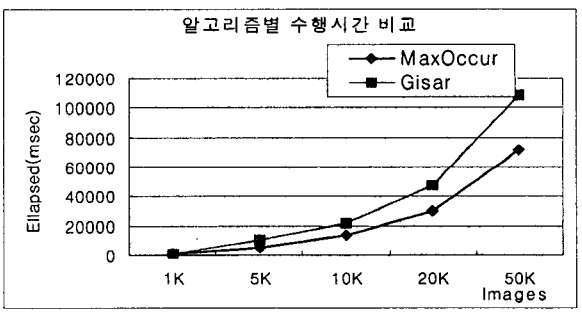
```
D : Database
F_n ← {MaxOccur 알고리즘의 빈발 항목 집합}

for(k ←2; F_k ≠ ∅; k++)
{
  SFC← gisar_generate(F_k, F_k)
  for all candidates sfc ∈ SFC
  {
    sfc.Ssup ← spatial_support(sfc)
    if(sfc.Ssup ≥ MSsup)
      SFS_k = { sfc ∈ SFC | sfc.Ssup ≥ MSsup }
  }
}

gisar_generate(P, Q)
{
  SELECT p.Item_1, q.Item_1, .... , q.Item_k-1
  FROM P p, Q q
  WHERE p.Item_1 <= q.Item_1, p.Item_2 <= q.Item_2, ... ,
        p.Item_k-1 <= q.Item_k-1;
}

spatial_support(sfc)
{
  sfc의 항목들간 거리가 same-cell 또는 adjacent개수 /
    sfc에서 k항목 부분 집합 수
}
```

(그림 2) GISAR 알고리즘

알고리즘 확장성과 성능을 비교하기 위해 각 이미지당 15개 정도의 오브젝트를 가진 가상 이미지를 설정하고 동일한 크기의 이미지 집합을 생성했다. 각 이미지에서 오브젝트들의 위치 좌표는 MBR의 중심점을 임의의 X, Y좌표로 결정하였다.

테스트 환경은 Pentium M 1.6, 1GB, Windows XP Professional, J2SDK 1.4.1 환경에서 수행하였고 랜덤하게 생성된 50,000건의 이미지를 사용하여 각 알고리즘의 성능 평가를 수행하였다.

MaxOccur 알고리즘은 오브젝트 기반의 지지도를 이용하였고 본 논문에서 제안한 알고리즘은 그 결과에 공간 지지도를 적용하였다.




(그림 3) 알고리즘별 수행 시간 비교

(그림 3)은 두 가지 알고리즘에 대한 평균 실행 시간을 보여 준다.

MaxOccur 알고리즘은 공간 지지도를 적용하여 최소 지지도는 0.1로, GISAR 알고리즘은 MaxOccur 알고리즘의 결과로 추출된 빈발 항목 집합들에 다시 공간 지지도 0.125를 적용하여 빈발 항목들에 대한 셀들 간의 위치적

인 관계를 이용하여 최소 공간 지지도를 만족하지 못하면 빈발 항목에서 제외된다.

GISAR 알고리즘은 MaxOccur 알고리즘의 결과를 후보 항목 집합으로 하여 *SF*를 적용한 결과이므로 속도와 마이닝 결과 성능은 감소하였지만 본 논문에서 제안하고자 했던 부분인 빈발 항목 집합들의 공간적인 관계를 살펴봄 으로써 이미지에서 빈발한 오브젝트들 간의 공간적인 관계를 살펴 볼 수 있다.

<표 3>은 이미지 개수별 각 알고리즘에서 찾아낸 연관 규칙 개수이다. 1,000개의 이미지에서 빈발 항목 항목이면서 공간 연관 규칙을 가지는 빈발 항목 집합은 130임을 알 수 있다.

<표 3> 알고리즘별 빈발 항목 집합 검색 결과

| # of images | MaxOccur | GISAR |
|---|---|---|
| 1K | 324 | 130 |
| 5K | 438 | 257 |
| 10K | 476 | 296 |
| 20K | 663 | 472 |
| 50K | 680 | 512 |

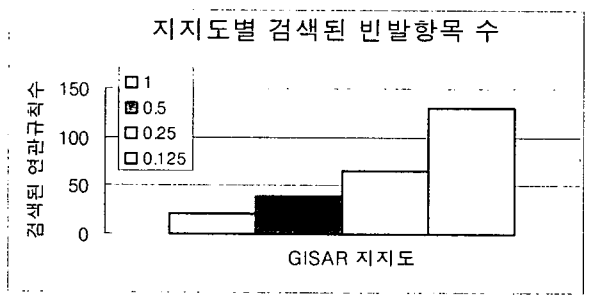
<표 4> GISAR알고리즘의 다른 지지도별 평균 수행 시간

(단위 msec)

| # of images | 0.125 | 0.25 | 0.5 | 1 |
|---|---|---|---|---|
| 1K | 971 | 1,111 | 541 | 360 |
| 5K | 10,815 | 8,712 | 6,159 | 5,047 |
| 10K | 21,831 | 18,457 | 11,456 | 9,073 |
| 20K | 47,098 | 35,511 | 21,321 | 15,722 |
| 50K | 108,596 | 77,351 | 51,814 | 42,851 |

<표 4>은 다른 지지도별 GISAR 알고리즘의 평균 수행 시간으로 지지도가 낮아질수록 수행 시간이 더 소요됨을 알 수 있다. 이와 같은 결과는 공간 지지도가 낮을수록 GISAR 알고리즘의 연산에 참여하게 되는 오브젝트의 수가 많아지기 때문이다.

(그림 4)는 GISAR 알고리즘 지지도별 빈발 항목 집합 검색 결과이다. 공간 지지도가 낮아질수록 빈발 항목 수도 많아지며 수행 시간도 길어짐을 알 수 있다.




(그림 4) GISAR알고리즘 지지도별 빈발 항목 집합 검색 결과

## 5. 결론 및 향후 연구

본 논문에서는 기존의 멀티미디어 연관 규칙 알고리즘인 Max occur 알고리즘에서 추출한 빈발 항목 집합의 결과들에 대하여 빈발 항목 집합들끼리의 공간적인 연관 관계를 고려하기 위하여 공간 분할 방법인 그리드 셀 기반으로 공간 분할 지수(spatial factor)인 SF를 이용한 이미지 공간 연관 규칙 마이닝 방법을 제시하였다. 새로 제안된 마이닝 기법에 최소 공간 지지도를 적용하여 항목들 간의 공간 관계를 살펴 봄 으로써 이미지 오브젝트들 간의 연관규칙을 마이닝 하는데 보다 유효한 알고리즘을 제안했다.

실험 결과 제안된 알고리즘은 기존의 이미지 연관 규칙 탐사 기법인 MaxOccur 보다 연관 규칙을 탐색하는데 있어서 공간적인 거리를 적용하였기 때문에 동일 시간 내 더 많은 수행시간이 소요 되었지만, 같은 종류의 이미지가 모여 있는 저장소에서 이미지 오브젝트 간의 공간적인 연관 관계를 발견하는 이미지 데이터 마이닝에 효과적이다.

공간 데이터 마이닝을 위한 기존의 알고리즘들은 대부분이 객체들 간의 거리 계산을 기반으로 하므로 데이터 양이 많아질수록 비용이 커지는 경우가 많이 발생한다. 또한 메모리 상주 데이터를 대상으로 하므로 대용량의 데이터인 경우에 효율이 떨어지는 문제점이 발생한다. 이러한 문제점들을 해결하기 위하여, 본 논문에서는 그리드 셀 구조를 기반으로 한 공간 연관 규칙 알고리즘을 제시하였다. 이 알고리즘에서는 기본적으로 셀 관련성을 기반으로 하여 실제 객체들 간의 거리 계산을 최소화함으로써 대용량의 공간 데이터에 대한 연관 규칙 마이닝에 효율적일 수 있다.

향후 연구 방향으로는 공간 데이터의 속성을 보다 잘 살릴 수 있는 공간 연관 규칙 탐사 기법의 연구 들이 필요하다.

## 참고문헌

[1] Krzysztof Koperski, Junas Adhikary, Jiawei Han "Spatial Data Mining: Progress and Challenges Survey Paper" SIGMOD'96 Workshop. on Research Issues on Data Mining and Knowledge Discovery (DMKD '96) 1996.

[2] Wei Lu, Jiawey Han, Beng Chin Ooi "Discovery of General Knowlege in Large spatial Databases" Proceedings of Far East Workshop on Geographic Information Systems. Singapore. 1993

[3] Krzystof Koperski, Jiawey Han "Discovery of spatial association rules in Geographic Information Databases" Advances in Spatial Databases. Proceedings of 4th Symposium, SSD'95. 1995

[4] 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, pp.461-470, March 2000.