밀도 기반 클러스터링을 이용한 효과적인 공간 특성화 방법의 설계 및 구현

°유재현*, 이주홍*, 전석주**, 박상호*
인하대학교 컴퓨터 정보공학과*, 서울교육대학교 컴퓨터 교육과**
you@datamining.inha.ac.kr*, juhong@inha.ac.kr*,
chunsj@snue.ac.kr**, parksangho@datamining.inha.ac.kr*

Design and Implementation of Effective Spatial Characterization using Density-Based Clustering

°Jae-Hyun You*, Ju-Hong Lee*, Seok-Ju Chun**, Sang-Ho Park*
Dept, of Computer Science & Information Engineering, Inha University*
Dept, of Computer Education, Seoul National University of Education**

요 약

최근 유비쿼터스 컴퓨팅의 관심이 증대되면서, 방대하고 다양한 형태의 데이터에 대한 효율성과 효과성을 고려한 지식 탐사방법연구의 필요성이 제기되었다. 기존의 지식 탐사방법에 대한 연구들은 방대한 비공간 데이터들의 지식을 효율적으로 탐사하고자 하였다. 그러나 기존의 연구는 탐사된 지식의 효율성만을 고려하여 유용한 지식탐사를 보장하지 못하는 문제점을 가진다. 따라서 본 논문은 공간 데이터 타입을 포함하는 대용량의 데이터들로부터 효과성을 보장하는 특성화 지식 탐사방법을 제안한다. 본 논문에서 제안하는 특성화 지식 탐사기법은 공간 및 비공간 데이터들의 특성을 나타내는 요약된 지식을 제공하며, 밀도 기반의 클러스터링 기법을 적용하여 특성화 지식 탐사의 효과성을 높인다.

1. 서 론

최근, 인간 생활의 편리성을 극대화하기 위한 유비쿼터스 컴퓨팅에 대한 연구가 활발히 진행되고 있다. 이러한 연구들은 다음과 같은 대부류로 나누어 생각할 수 있다. 첫째, 유비쿼터스 환경은 분산 네트워크상의 방대한 양의 데이터를 처리해야 하며, 처리되는 데이터의 형태는 기존의 데이터 형태를 포함하는 더욱 다양한 형태로 구성되므로, 그에 대한 연구가 요구된다. 예를 들어, 상황 인식을 위한 인간의 정적 및 동적 위치 정보와 다양한 센서로부터 입력된 온도, 습도 등 다양한 형태의 데이터들을 처리해야 한다. 둘째, 질 높은 서비스를 인간에게 제공하기 위해서는 유비쿼터스 컴퓨팅의 과정에서 얻어진 지식이 인간 개인에게 매우 유용하여야 한다. 이러한 효과성이 보장된 다양한 지식은 기존의 데이터 마이닝의 기법들을 통하여 탐사될 수 있으며, 처리되는 데이터의 형태가 다양해짐에 따라서, 기존 데이터마이닝 기법들도 확장 및 연구되어야 한다.

기존의 데이터마이닝 기법들은 연관, 분류, 군집, 경향, 특성화 형태의 지식들을 탐사할 수 있다. 탐사된 지식들은 비공간 데이터들 사이의 본질적인 관계를 찾아내며, 축약된 방법으로 데이터의 규칙성을 찾아내는 공통점을 가지는 반면, 찾아진 규칙은 각각 다른 형태로 표현되어지며, 서로 독립적인 의미를 지닌다. 기존 데이터마이닝을 확장시킨 공간 데이터마이닝은 공간적 속성이 탐사된 규칙에 중요한 요소로 작용하여, 탐사된 규칙을 확장시킨다. 즉, 기존 데이터마이닝의 특성화 규칙은 데이터베이스(D)에서 "A속성=α AND B속성=β AND C속성=*" 형태의 특성화 규칙을 발견한다. 여기에서, A, B, C는 공간적 속성과 비관련성을 지닌다. 그에 반해서, 공간 데이터마이닝은 공간 데이터베이스(S)에서 위와 동일한 특성화 규칙 "A속성=α AND B속성=β AND C속성=*"을 생성하나, 이 때, A, B, C는 공간 객체와 밀접한 관련성을 가진다. 즉, 공간 데이터마이닝은 기존 데이터마이닝 기법과 달리 공간적 속성을 고려한 확장된 지식을 발견한다. 이러한 기존 데이터마이닝 기법의 공간 데이터마이닝으로의 확장에 대한 연구는 유비쿼터스 환경 실현을 위해서 필요한 연구 분야이다.

따라서, 본 논문은 기존의 특성화기법을 공간 특성화 기법으로 확장하여 특성화 규칙을 생성하고, 밀도 기반의 클러스터링기법을 적용하여 생성된 규칙의 효과성을 높이고자 한다.

1) 본 연구는 대한 IT연구센터 육성·지원사업의 연구결과로 수행되었음

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 관련 연구로 기존의 공간 데이터마이닝 시스템과 기존 특성화 기법에 대해 살펴본다. 3장에서는 본 논문에서 제안하는 공간 데이터베이스 기반의 공간 특성화를 위한 시스템을 제안하고, 4장에서는 밀도 기반 클러스터링을 적용한 공간 특성화기법을 제안한다. 마지막으로 5장에서는 결론에 대하여 언급한다.

2. 관련 연구

공간 지식탐사를 지원하는 대표적 공간 데이터마이닝 시스템에는 GeoMiner와 Economic Geography 등이 있다.

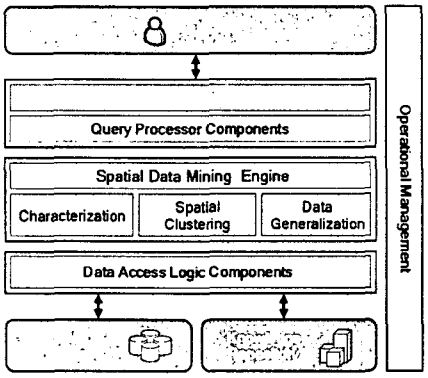
Lu&Han의 GeoMiner시스템은 비공간 데이터마이닝을 위한 DBMiner 시스템을 공간과 비공간 데이터마이닝을 위해 확장한 시스템이다[5]. GeoMiner시스템은 DBMiner에서 지원하는 연관, 군집, 특성화, 분류의 지식형태를 확장시켜, 공간 연관, 공간 군집, 공간 특성화, 공간 분류의 지식탐사를 가능하게 확장하였다. GeoMiner는 공간 특성화를 위해서 Non-Spatial Data Dominant Generalization(NSD)과 Spatial Data Dominant Generalization(SD) 알고리즘을 제안하였다. NSD알고리즘은 비공간 속성의 일반화 과정과 공간 속성을 고려한 합병 순으로 지식 탐사 과정을 수행하며, SD알고리즘은 공간 속성의 일반화 과정과 비공간 속성의 합병의 순으로 특성화를 수행한다. 이 때, NSD와 SD에 공통적으로 수행되는 일반화 작업은 각각 비공간과 공간 속성의 개념 계층을 필요로 하게 된다. 이러한 개념 계층은 일반화의 집계연산을 가능하게 하며, 구성된 개념 계층에 따라서 다양한 형태의 지식이 탐사될 수 있다. 이러한 GeoMiner의 장점은 사용자가 제공하는 임계치에 의해서, 탐사된 지식의 형태를 다양하게 변화시킬 수 있는 장점을 가진다. 그러나 일반화를 위한 개념 계층이 필요 영역의 전문가들에 의하여 각각 구성되어져야 하는 작업이 요구되는 단점을 가진다.

Ester는 BAVARIA 데이터베이스를 기반으로 Economic Geography시스템을 제안하였다[2]. 이 시스템은 전체 데이터베이스를 이용하지 않고, 공간과 비공간의 상대 빈도수 특성들을 이용하여 특성화를 수행한다. 즉, 주어진 목표 지점과 인접한 공간 및 비공간 객체들의 상대 빈도수만을 이용하여 특성화의 효율을 높였다. 이러한 Economic Geography는 주어진 목표 영역을 사용자에게 선택할 기회를 제공하여 보다 쉽고 편리하게 지식을 탐사할 수 있는 장점을 가진다. 그러나 목표 지점선택의 어려움과 목표 지점에 이웃한 영역의 상대 빈도수가 목표 지점의 상대 빈도수보다

높은 경우 필요이상의 이웃 확장성을 가지며, 이는 예상된 결과보다 낮은 흥미도를 가지는 지식을 탐사하게 되는 문제점을 가진다.

3. 공간 특성화를 위한 마이닝 시스템의 구조 및 설계

본 논문에서 제안하는 밀도 기반 클러스터링을 적용한 공간 특성화를 위한 마이닝 시스템은 다음 [그림 1]과 같다. 공간 특성화를 위한 마이닝 시스템은 다음과 같은 구성 요소를 가진다. 공간 특성화를 지원하는 데이터마이닝 질의 처리기[4], 공간 특성화 엔진, 공간 데이터베이스, 공간 특성화를 지원하는 개념 계층 저장소이다.



[그림 3] 공간 특성화 마이닝 시스템

이러한 공간 특성화를 위한 마이닝 시스템은 다음과 같은 특성을 지닌다. 첫째, 질의 처리기는 공간 데이터마이닝의 특성화, 연관 규칙, 분류, 경향등을 지원한다. 둘째, 본 시스템에서 이용하는 공간 데이터베이스는 GMS를 이용한다. 셋째, 공간 특성화를 지원하는 개념 계층에 대한 인터페이스를 통하여, 사용자가 직접 개념 계층을 구성할 수 있다.

3.1 공간 데이터마이닝을 위한 질의 처리기

공간 데이터마이닝을 위한 질의처리기는 본 논문의 사전 연구 결과인 SMQL을 이용하였다[4]. SMQL은 공간 마이닝을 위한 질의어 규격을 정의하였고, 각 기능을 모듈별로 구성하여 연관, 분류, 군집, 경향, 특성화를 지원하는 질의 처리기이다.

3.2 공간 특성화 엔진

공간 특성화 엔진은 공간 특성화, 밀도 기반 클러스터링, 데이터 일반화의 3가지 모듈로 구성된다. 공간 특성화 모듈은 공간 데이터베이스로부터 입력된 데이터들의 튜플단위의 집계 연산을 수행하며, 밀도 기반 클러스터링모듈은 공간적 속성의 밀도를 기반으로 공간 객체를 군집화하는 모듈이다. 또한, 데이터 일반화 모듈은 공간 특성화 작업을 수행하는 중에 필요한 데이터 변환작업을 수행하는 모듈이다.

3.3 공간 데이터베이스

본 시스템은 GMS 공간 데이터베이스를 이용한다[6]. GMS는 SQL92 표준을 기반으로 OGC에서 표준으로 제안하는 7개의 기본 공간 데이터 타입과 9개의 공간 관계, 확장된 공간 데이터 타입 및 공간 관계연산자 및 공간 함수를 지원하는 공간 데이터베이스이다.

3.4 공간 특성화를 지원하는 개념 계층 저장소

공간 특성화를 지원하는 개념 계층 저장소는 속성의 변환작업(즉, 일반화 작업)에 필요한 대치 속성값을 저장하며, 이러한 대치 속성값들은 공간 특성화 엔진의 공간 특성화 과정 중 데이터 일반화의 모듈 수행시에 적절히 이용된다. 또한, 이러한 개념 계층의 구성을 시스템 사용자나 지식 전문가에 의해 제공되어 질 수 있도록 공간 마이닝을 위한 질의 처리기와 함께 운용된다.

4. 밀도 기반 클러스터링을 적용한 공간 특성화

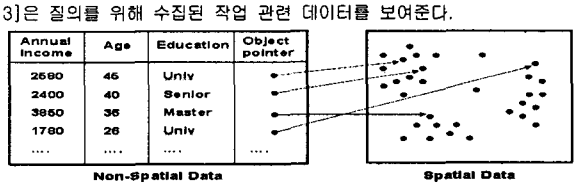
본 논문에서 제안하는 밀도 기반 클러스터링을 적용한 공간 특성화 과정은 위의 공간 특성화 엔진에서 3개의 모듈에 의해서 수행되며, 크게 공간 특성화와 밀도 기반의 클러스터링 부분으로 나누어지며, 세부적으로 총 5단계의 과정을 통하여 수행된다. 제안한 공간 특성화 과정은 [그림 2]와 같다.

입력 : 공간 및 비공간 속성을 포함하는 공간 데이터,
공간 및 비공간 개념 계층, 사용자 임계치
출력 : 공간 특성화 규칙
방법 :

1. 사용자의 질의에 의한 특성화 관련 공간, 비공간 데이터 수집한다.
2. 사용자 임계치를 만족 할 때까지 공간 및 비공간 데이터에 대한 일반화를 수행한다.
   (1) 작업데이터의 불필요한 속성을 제거한다.
   (2) 공간 및 비공간 개념계층이 존재한다면, 데이터를 상위 레벨로 일반화를 수행한다.
3. 단계 2의 공간 객체를 대상으로 밀도 기반 클러스터링을 수행한다.
4. 얻어진 데이터에 대한 집계연산을 수행한다.
5. 결과로부터 일반화된 규칙이나 패턴을 찾는다.

[그림 2] 밀도 기반 클러스터링을 적용한 공간 특성화

위의 [그림 2]의 공간 특성화 방법에 대한 예는 질의 1.과 같고 [그림 3]은 질의를 위해 수집된 작업 관련 데이터를 보여준다.



[그림 3] 작업 관련 데이터 수집

질의 1 : 인천 지역에 대한 여성 거주자를 대상으로 소득, 학력과 나이에 대한 특성화를 수행하시오. SMQL기반의 질의문은 아래와 같다[4].

```
MINE Characteristic as woman_pattern
USING HIERARCHY H_inome, H_education, H_age
USING Clustering distance 30
for annual_income, education, age
from census
where province = "인천", gender = "F";
set distinct_value threshold 20
```

질의문에 대한 공간 특성화의 수행방법은 아래의 4단계로 구성된다.

[단계1] 사용자에 의해 주어진 위와 같은 질의문은 질의처리기(SMQL)의 파싱 과정을 통하여 토큰을 생성하고, 생성된 토큰을 각 모듈과 연결하여 작업 관련 공간 데이터 수집 과정을 수행한다. 위의 경우, census 테이블의 annual_income, education, age, 그리고 공간 속성값들이 튜플 단위로 수집된다. 여기에서, 각 튜플은 하나의 공간 객체를 의미한다.

[단계 2] 이 단계에서는 수집된 공간 객체들의 속성값들의 일반화 작업을 수행한다. 일반화 작업은 각 속성마다 구성된 개념 계층을 이용하여, 기존의 속성값을 상위 개념의 속성값에 대한 대치 작업(즉, 일반화 과정)을 통하여 수행된다. 이러한 대치작업은 공통된 속성값들을 가지는 작업 관련 데이터의 수와 초기 임계치와의 비교를 통하여, 임계치보다 작은 공통된 속성값들을 가지는 튜플을 제거하고, 한번의 집계 연산을 수행한다. 한번의 수행된 집계 연산의 결과를 다시 사용자가 주어진 임계치보다 작다면 개념 계층을 통하여, 속성값을 상위 개념의 속성값으로 대치하고, 이러한 과정은 임계치를 만족할 때까지 반복적으로 수행한다.

[그림 4] 연령에 대한 개념 계층

[그림 4]는 상위 개념의 속성값으로 대치과정을 위한 개념 계층을 보여준다. <표 1>은 [그림 4]의 개념 계층을 이용하여 [그림 3]의 작업

<표 3> 개념 계층을 적용한 일반화 결과

Annual_Income	Age	Education	Object Pointer
middle	middle age	higher	•
middle	middle age	secondary	•
middle-high	youth	higher	•
low-middle	youth	higher	•
...	...	...	...

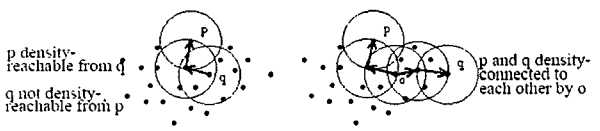
관련 데이터의 일반화된 결과를 보여준다. 이렇게 대치된 결과는 공통된 속성값을 가지는 튜플을 생성하게 되므로, 집계 연산을 가능하게 한다.

[단계 3] 이 단계에서는 일반화 과정을 거친 결과 튜플들의 공간 속성들을 이용한 군집화 작업을 수행한다. 위의 <표 1>의 Object Pointer속성은 각 공간 객체의 공간 정보를 나타내는 POINTER를 저장하며, POINTER들의 각 공간 속성(예를 들어, 공간 좌표)을 이용하여, 공간 군집화를 수행한다. 본 논문에서 사용한 군집화는 대표적인 밀도 기반 클러스터링 방법인 DBSCAN을 기반으로 한다. 그러나 DBSCAN은 단지 Pointer만을 이용한 군집화를 수행한다. 본 공간 특성화 시스템에서 이용하는 공간 데이터베이스는 POINTER 데이터 타입뿐만 아니라, POLYGON 타입의 공간 객체를 지원하므로, 본 논문에서는 POLYGON과 POINTER를 지원하게 DBSCAN을 수정하였다. 수정된 군집화방법은 주어진 반지름값(ε)을 가지는 클러스터 중 최소한 사용자 정의 임계치인 MinPtr이상의 공간 객체를 포함하는 클러스터를 선택하고, 선택된 클러스터 안의 공간 객체를 중심으로 하위 클러스터를 생성하여, 하위의 클러스터들을 병합하여 하나의 하위 클러스터를 포함하는 클러스터를 생성한다. 즉, 다음과 같은 두 가지 조건을 만족하는 공간 클러스터들을 발견하여 병합한다.

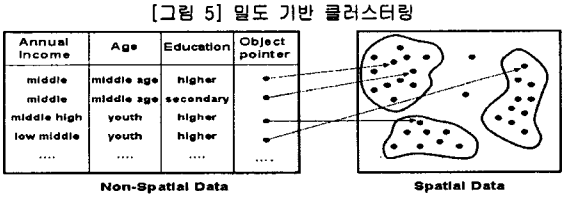
조건_1 : $Eps\text{-}neighborhood(N_{Eps}(p))=\{q \in D \mid dist(p,q) \le Eps\}$

조건_2 : $|N_{Eps}(p)| \ge MinPtr$.

[그림 5]은 본 논문에서 제안한 밀도 기반 클러스터링의 병합과정을, [그림 6]은 클러스터링을 적용한 공간 특성화의 결과를 보여준다.



[그림 5] 밀도 기반 클러스터링



[그림 6] 클러스터링을 적용한 특성화 결과

[단계 4] 본 단계에서는 위의 3단계의 결과를 이용하여 튜플병합과 집계연산을 수행한다. 그리고 최종적으로 사용자 임계치와 비교하여, 임계치보다 작은 지식을 가지면, 각 튜플의 속성들에 대한 정보 이득(Information Gain)을 이용하여 지식의 흥미도를 측정한다. 하나의 속성에 대한 데이터의 집합 S에 대한 엔트로피와 속성이 가지는 가중치 이용한 정보이득은 다음과 같이 정의 된다.

$$Gain(S) = E - \{ W_a E_a + W_b E_b \}$$

질의 2 : "인천지역에서 나이가 30대인 거주자를 대상으로 신용불량자와 지출에 대한 특성화를 수행하여라." 라는 질의문은 아래와 같다.

```
MINE characteristic as credit_pattern
USING HIERARCHY H_inome, H_expenses
USING Clustering distance 20
for credit_defaulter, expenses
from census
where province = "인천", age<29 and age>40
set distinct_value threshold 20
```

<표 2> 질의 2.에 대한 공간 특성화

cluster	address	credit defaulter	expenses	aggregation	Object Pointer
C1	Nam-gu	No	Low	197	•
C1	Nam-gu	Yes	Low	42	•
...	...	...	...	..	

(a) 지도상의 공간객체 (b) 밀도 기반 클러스터링

[그림 7] 주어진 영역에 대한 밀도 기반의 클러스터링

위의 질의 2.는 먼저 질의에 맞는 작업데이터를 데이터베이스로부터 가져와 일반화 작업 후 [그림 7]의 (a)와 같이 인천지역의 모든 객체들을 대상으로 수행하는 것이 아니라 밀도 기반의 클러스터링을 이용하여 공간객체가 많이 모여 있는 밀도가 높은 지역을 대상으로 수행한다. 즉, 사용자가 지정한 거리 사이에 떨어져 있는 공간 객체들을 각각의 클러스터로 만듦으로서, [그림 7]의 (b)와 같은 밀도 기반의 클러스터를 얻게 된다. 클러스터링을 통해 얻어진 공간 영역에 대해 공간 특성화를 수행하면 <표 2>와 같은 결과를 얻게 된다. 이러한 결과는 각 클러스터의 공간정보와 비 공간정보를 사용자에게 제공한다. 따라서 사용자는 기존의 공간 특성화에 비해 분석하고자하는 지역에 대한 보다 자세한 지식 탐사를 수행할 수 있다.

5. 결론

최근 지리정보시스템과 같은 방대한 양의 공간 데이터를 다루는 응용시스템에서 공간 데이터베이스로부터 규칙적인 특성, 혹은 흥미로운 지식을 추출해내는 공간 데이터마이닝에 대한 중요성이 높아지고 있다. 그러나 기존의 공간 특성화방법들은 주어진 공간영역에 대한 효과적인 지식탐사를 하는데 있어서, 사용자가 미리 정의한 지리영역에 크게 영향을 받아 제한된 영역에 대한 특성화를 수행하게 된다. 본 논문은 밀도 기반 클러스터링을 적용한 특성화 방법을 제안하고, 제안된 특성화 방법을 적용하여 새로운 공간 데이터마이닝 시스템을 설계하였다. 제안된 밀도 기반의 클러스터링을 적용한 특성화 방법은 공간 및 비공간 데이터에 대하여 기존의 방법보다 탐사된 지식의 효과성을 향상시켰다.

참 고 문 헌

[1] J. Han and Y. Fu : Dynamic Generation and Refinement of Concept hierarcies for Konwledge Discovery in Databases, AAAI'94 Workshop on Knowledge Discovery in Databases(KDD'94), Seattle, p.157-168, WA, July (1994)

[2] Ester, M., Kriegel, H.-P.and Sander, J. : Algorithms and applications for spatial data mining, in H. J. Miller and J. Han (eds.) Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, p160-187., (2001)

[3] Ester M., Kriegel H.-P., Sander J., Xu X. : A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proc. 2nd int. Conf. on Knowledge Discovery and Data Mining (KDD '96), Portland, (1996)

[4] 박선, 박상호, 안찬민, 이윤석, 이주홍 : SIMS를 위한 공간 데이터 마이닝 질의 언어, 한국정보과학회 춘계 학술발표논문집 제 31 권 제 1 호, p70-72, (2003)

[5] J. Han, K.Koperski and N. Stefanovic : GeoMiner : A system prototype foe spatial data mining, Proceedings of 1997 ACM-SIGMOD International Conference on Management of Data(SIGMOD '97), p553-556, (1997)

[6] 박상근, 박순영, 정원일, 김명근, 배해영 : GMS : 공간 데이터베이스 관리 시스템, 공동 춘계학술대회, p217-224, (2003)