韓國 컴퓨터情報學會 論文誌
第11卷 第2號, 2006. 5.

2006-11-2-1-6

밀도 클러스터링을 이용한 공간 특성화 시스템 설계 및 구현

유 재 현*, 이 주 홍**, 박 태 수*, 안 찬 민*, 박 상 호*, 홍 준 식***

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

Jae Hyun You*, Ju Hong Lee**, Tae Su Park*, Chan Min Ahn*, Sang Ho Park*, Jun Sik Hong***

요 약

최근 유비쿼터스 컴퓨팅의 관심이 증대되면서, 방대하고 다양한 형태의 데이터에 대한 효율성과 효과성을 고려한 지식 탐사연구의 필요성이 요구된다. 공간 특성화 방법은 공간과 비공간 속성들을 고려하여 특성화 지식을 발견하는 방법으로, 기존의 특성화 방법을 확장하여 공간 영역에 대한 다양한 형태의 지식을 발견할 수 있다. 기존 공간 특성화기법에 대한 연구들은 다음과 같은 문제점을 가진다. 첫째, 기존의 연구는 탐사된 지식의 결과가 다각적인 공간 분석을 수행하지 못하는 문제점을 가진다. 둘째, 공간 탐색 시 사용자에 의해 미리 정해진 위치 영역만을 고려하여 탐색함으로 유용한 지식탐사를 보장하지 못하는 문제점을 가진다. 따라서 본 연구에서는 밀도 기반의 클러스터링이 적용된 새로운 공간 특성화기법을 제안한다.

Abstract

LRecently, with increasing interest in ubiquitous computing, knowledge discovery method is needed with consideration of the efficiency and the effectiveness of wide range and various forms of data. Spatial Characterization which extends former characterization method with consideration of spatial and non-spatial property enables to find various form of knowledge in spatial region. The previous spatial characterization methods have the problems as follows. Firstly, former study shows the problem that the result of searched knowledge is unable to perform the multiple spatial analysis. Secondly, it is unable to secure the useful knowledge search since it searches the limited spatial region which is allocated by the user. Thus, this study suggests spatial characterization which applies to density based clustering.

▸ Keyword : 공간 데이터 마이닝(Spatial Data Mining), 공간 특성화(Spatial Characterization), 지리 정보 시스템(Geographic Information System), 데이터 마이닝(Data Mining)

• 제1저자 : 유재현 • 교신저자 : 이주홍
• 접수일 : 2006.02.16 심사완료일 : 2006.05.18
* 인하대학교 컴퓨터정보공학과, ** 인하대학교 컴퓨터공학부 부교수, 교신저자,
*** 영동대학교 전자의용공학부 전임강사
◆ 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성.지원사업의 연구결과로 수행되었음.
◆ 이 논문은 2006년도 두뇌한국21사업에 의하여 지원되었음

I. 서 론

최근 공간 데이터를 다루는 다양한 종류의 응용시스템에서 공간 데이터베이스로부터 규칙적인 특성, 혹은 흥미로운 지식을 발견하고자하는 연구에 대한 중요성이 높아지고 있다. 기존의 데이터마이닝 기법들은 관계형 데이터베이스를 기반으로 연관, 분류, 군집, 경향, 특성화 등의 다양한 기법들에 의하여 주어진 데이터로부터 지식들을 탐사할 수 있다[1]. 기존 데이터마이닝 개념을 확장시킨 공간 데이터마이닝은 위성정보나 지리정보시스템(GIS)의 데이터로부터 공간적 속성을 가진 탐사된 규칙에 중요한 요소로 작용하여 방대한 지리 정보로부터 발견된 규칙을 확장한다[2,3].

그러나 기존 공간 특성화 시스템은 다음과 같은 문제점을 가진다[4,5,6]. 첫째, 현존하는 공간 특성화 방법은 사용자가 분석하고자하는 공간 영역에 대한 분석을 원활하게 수행하지 못한다. 일반화기반의 공간 특성화방법은 사용자가 질의를 할 경우, 질의 영역에 포함되는 공간객체에 대한 일반화 작업을 수행한다. 이러한 공간 객체에 대한 일반화 수행 시, 영역 전문가 혹은 도메인 전문가에 의해 미리 정의해 놓은 영역만을 사용함으로 사용자가 분석하고자하는 공간영역의 범위를 줄이게 된다. 그러나 "인구가 밀집한 어떤 지역을 대상으로 소비와 수입에 대한 패턴을 분석하여라."와 같은 유형의 질의는 불필요한 지식이 포함되어 있는 지역이 제외된 공간 영역만을 대상으로 분석하며, 이러한 결과는 사용자에게 공간 영역에 대해 추론할 수 있는 범위를 넓혀주고 다양한 질의를 제공할 수 있게 된다. 둘째, 공간과 비공간 속성을 이용한 특성화에 의한 결과를 효과적으로 유도해내기 위해서는 사용자가 공간 특성화 과정에 포함되는 공간 속성 범위를 직접 질의에 포함시켜 주어야 한다. 이러한 사용자에 의한 공간적 영역의 지정은 사용자에게 주어진 공간 영역에 대한 다각적인 이해와 세분화된 공간 영역으로부터 흥미로운 패턴과 지식탐사를 어렵게 한다는 문제점을 가진다.

따라서 이러한 문제점을 해결하기 위하여 본 연구에서는 일반화기법을 확장한 공간 데이터마이닝 모듈과 밀도 기반의 클러스터링 모듈을 통합한 공간 데이터마이닝 시스템을 제안한다. 제안된 방법은 기존의 특성화기법을 공간 특성화 기법으로 확장하여 특성화 규칙을 생성하고, 밀도 기반의 클러스터링 기법을 적용하여 생성된 규칙의 효과성을 높이고자 한다[7].

본 연구의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 관련 연구로 기존의 공간 데이터마이닝 시스템과 기존 특성화 기법에 대해 살펴본다. 3장에서는 본 연구에서 제안하는 공간 데이터베이스 기반의 공간 특성화를 위한 시스템의 구성요소에 대한 설명을 하고, 4장에서는 확장된 공간 데이터마이닝을 위하여 밀도 기반 클러스터링을 적용한 공간 특성화기법을 제안한다. 마지막으로 5장에서는 결론 및 향후 연구방향에 대해서 기술한다.

II. 관련 연구

최근까지의 공간 데이터마이닝의 기법에 관한 연구는 다음과 같다. 공간 및 비공간 데이터의 일반화된 요약정보를 추출하는 공간 특성화기법(Spatial Characterization)과 데이터간의 연관관계를 발견하는 공간 연관 규칙기법(Spatial Association rule), 공간 객체들을 연관된 그룹으로 군집화하는 공간 클러스터링 기법(Spatial Clustering) 등이 대표적인 기법으로 분류된다[1,4,7,9]. 이러한 기법을 사용하여 공간 및 비공간 데이터로부터 지식을 발견하는 공간 데이터마이닝 과정은 <그림 1>과 같다[1].

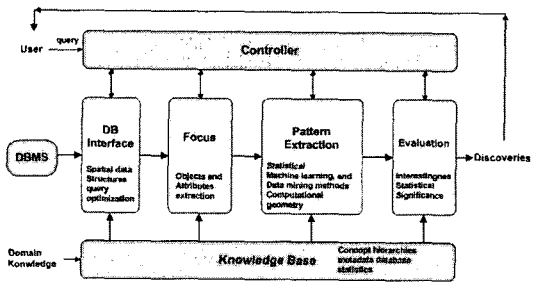


그림 1. 공간 데이터마이닝의 과정
Fig. 1 A process of spatial data mining

대표적 공간 데이터마이닝 시스템에는 GeoMiner와 Economic Geography 등이 있다. Lu&Han의 GeoMiner시스템은 비공간 데이터마이닝을 위한 DBMiner시스템을 공간과 비공간 데이터마이닝을 위해 확장한 시스템이다[6].

GeoMiner시스템은 DBMiner에서 지원하는 연관, 군집, 특성화, 분류의 지식형태를 확장시켜 공간 지식탐사를 가능하게 한다. GeoMiner는 공간 특성화를 위해서 NSD(Non-Spatial Data Dominant Generalization)과 SD(Spatial Data Dominant Generalization) 알고리즘을 제안하였다. NSD알고리즘은 비공간 속성의 일반화 과정과 공간 속성을 고려한 합병 순으로 지식 탐사과정을 수행하며, SD알고리즘은 공간 속성의 일반화 과정과 비공간 속성의 합병의 순으로 특성화를 수행한다. GeoMiner의 장점은 사용자가 제공하는 임계값에 의해서, 탐사된 지식의 형태를 다양하게 변화시킬 수 있는 장점을 가진다. 그러나 일반화를 위한 개념 계층이 전문가들에 의하여 구성되어져야 하며, 사용자에게 다각적인 분석을 제공하지 못하는 단점을 가진다. Ester는 BAVARIA 데이터베이스를 기반으로 Economic Geography시스템을 제안하였다[7]. 이 시스템은 전체 데이터베이스를 이용하지 않고, 공간과 비공간의 상대 빈도수 특성들을 이용하여 특성화를 수행한다. 즉, 주어진 목표 지점과 목표 지점에 인접한 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.

III. 밀도 기반 클러스터링을 이용한 공간 특성화 방법

이 장에서는 기존 공간 데이터마이닝 시스템의 문제점을 해결하는 방안으로, 밀도 기반의 클러스터링을 이용한 공간 특성화 방법에 대하여 기술한다. 본 논문에서 제안하는 방법은 첫째, 밀도 기반 클러스터링을 특성화 과정에 적용하고, 공간 속성을 가진 데이터에 대한 사전작업을 수행하여 주어진 공간 영역에서 발견된 특성화 지식의 효과성을 높인다. 둘째, 공간 계념 계층을 이용한 자동 공간 속성을 질의에 포함시켜 사용자에게 다양한 분석의 결과와 질의 편리성을 제공하는 장점을 가진다. 다음은 본 연구에서 사용하는 기법인 밀도 기반 클러스터링을 적용한 공간 특성화 기법에 대하여 기술한다.

3.1 제안하는 공간 특성화방법

공간 특성화(spatial characterization)는 공간상에 주어진 공간객체와 각 객체가 내포하고 있는 데이터 집합으로부터 탐색하고자하는 공간 영역에 대한 데이터 클래스의 전체적인 윤곽을 파악하는 방법으로, 사용자에게 간략하고 간결한 요약정보를 제공한다. 본 연구에서 제안하는 밀도 기반 클러스터링을 이용한 공간 특성화 시스템은 세부적으로 총 5단계의 과정을 통하여 수행된다. 제안하는 공간 특성화의 수행과정은 〈그림 2〉과 같다.

사용자가 지식이나 패턴을 발견하고자하는 영역범위에 대한 공간 및 비공간 데이터를 공간 데이터베이스로부터 수집한다. 이렇게 수집된 데이터에 대하여 공간 특성화방법을 적용한다. 공간 특성화는 공간 및 비공간 데이터를 일반화하는 작업과 관련이 깊다. 제안하는 공간 특성화 방법에서는 두 가지 데이터 일반화로 속성제거 일반화와 개념 계층을 사용하는 일반화방법으로 수행된다[5,8].

먼저 속성제거 일반화방법은 데이터베이스로부터 수집된 작업관련 비공간 데이터를 요약정보로 변환하는 과정에서 사용된다.

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 과정
Fig. 2 A process of spatial characterization using density-based clustering

이 방법은 작업관련데이터를 요약정보로 변환하는 과정에서 요약이 되지 않는 데이터를 특성화 기법에 불필요한 데이터로 판명하여 제거한다. 속성제거는 다음의 규칙에 따라 수행되며, 〈알고리즘 1〉과 같다. 속성제거는 하나의 속성이 가지는 데이터가 사용자가 지정한 임계값보다 많은 경

우와 입력된 데이터가 일반화를 위한 상위단계의 개념 계층이 없는 경우에 속성제거가 된다.

알고리즘 1. 속성제거 알고리즘
Algorithm. 1 Attribute removal algorithm

```
Precondition
: make the task-relevant data from SDB
Input Parameter
:TaskRelvantData, SizeOfFeild, SizeOfRow

Attribute_removal(TaskRelvantData, SizeOfFeild,
                  SizeOfRow)
FOR I FROM 1 TO SizeOfFeild DO
FOR J FROM 1 TO SizeOfRow DO
CountOfDistinctTuple := TaskRelvantData.get(i,j);
IF CountOfDistinctTuple >= UserThreshold THEN
REMOVE Attribute;
ELSE
Using Attribute;
END IF
END FOR
END FOR
END;
```

속성제거 일반화단계를 거치면 개념 계층을 이용한 일반화를 수행한다. 데이터베이스에 저장되어 있는 데이터는 개념 계층을 이용하여 요약된 정보로 바꾸어 줄 수 있다. 상위 개념 단계에서는 하위 개념 단계에서보다 데이터와 객체들의 집합들이 더 일반적인 정보를 가지고 있으며, 더 요약한 표현으로 나타낸다. 즉, 하위 개념 단계로부터 상위 개념 단계로 올라가면서 적절하게 데이터집합을 추상화하는 것이다. 그렇게 일반화된 데이터에서 지식을 추출하게 된다. 이러한 일반화기법은 공간 특성화방법에 중추적인 역할을 한다.

공간 특성화기법을 사용하여 일반화된 지식을 공간 데이터베이스로부터 추출하기 위해서는 공간 데이터와 비공간 데이터에 대한 일반화가 모두 요구된다. 따라서 데이터 일반화 모듈에서는 공간, 비공간 속성에 대한 일반화 작업을 수행하기 위하여 지식저장소에 미리 저장되어 있는 개념계층을 사용하거나 혹은 본 시스템에서 제공하는 사용자 정의 개념 계층을 이용하여 사용자가 직접 개념계층을 정의하여 일반화 작업에서 사용한다.

3.2 밀도 기반 공간 클러스터링

공간 데이터 마이닝기법들은 공간데이터베이스에 저장된 비공간 데이터와 공간 데이터를 사용한다. 공간 데이터는 공간상에 존재하는 공간 객체(또는, 공간 오브젝트)로서 위치속성을 나타내는 좌표로 표현된다. 각각의 공간 객체들은 관련된 비공간 데이터를 포함하고 있으며, 공간과 비공간 데이터는 서로 링크를 통해 연결된다. 공간 클러스터링은 사용자가 탐색하고자하는 영역에 대해 군집화를 수행한다. 본 연구에서 공간 클러스터링은 공간 특성화방법에서 사용되는 공간영역을 군집화하고 공간을 분석하기 위한 기본 도구로서 사용된다[7].

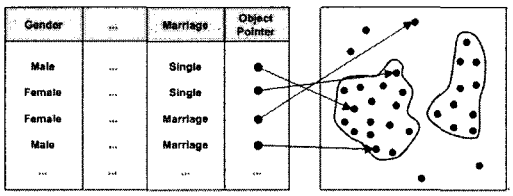


그림 3. 밀도 기반 클러스터링을 이용한 공간 특성화 결과
Fig. 3 A result of spatial characterization using density-based clustering

따라서 밀도 기반의 클러스터링은 데이터 일반화작업을 거친 공간 속성들을 이용하여 군집화 작업을 수행한다. 〈그림 3〉은 일반화 작업을 거친 공간과 비공간 데이터를 보여준다. 그 중 객체 포인터(Object Pointer)는 공간 속성으로 공간 데이터를 내포하고 있다. 공간 객체는 점(point) 또는 다각형(polygon)으로 구성되며 비공간 속성들과 함께 위치정보가 링크 형태로 연결되어 저장되어 있다. 공간 클러스터링은 이러한 각각의 공간 객체들의 위치속성을 이용하여 공간 군집화를 수행한다.

본 연구에서 사용한 군집화는 대표적인 밀도 기반 클러스터링 방법인 DBSCAN을 기반으로 한다. 그러나 공간상에서 좀 더 흥미로운 지식이나 패턴을 발견하기 위해서는 공간 객체를 표현하는 단순한 점 형태뿐만 아니라, 다각형 형태의 공간 객체를 다루어야 한다.

본 공간 특성화 시스템에서 이용하는 공간 데이터베이스는 점 데이터 타입은 물론 다각형 타입의 공간 객체를 지원하므로, 본 연구의 DBSCAN 기반의 클러스터링은 점 및 다각형 데이터도 지원한다. DBSCAN 방법을 사용함으로서 공간 영역에 대해 추론할 수 있는 범위를 넓혀주어 사용자에게 다각적인 분석을 하도록 유도한다.

DBSCAN은 밀도에 기초한 군집화 기법이다. 알고리즘은 충분하게 밀도가 높은 지역을 군집으로 키우고, 잡음값을 가진 공간적인 영역을 데이터베이스에서 임의의 형태인 군집을 찾는다. 밀도 기반의 방법은 주어진 반지름을 중심으로 포함되는 객체가 밀집한 부분을 군집하여 밀도 연결점의 최대화 집합으로 정의한다. 한 점 p의 E-neighborhood는 p로부터 반경 E내에 있는 이웃(neighborhood)의 집합이다.

$$\text{즉, } N_E(P) = \{q \in D | dist(p,q) \leq E\} \text{ 이다.}$$

밀도 기반의 클러스터링이 수행되는 방법은 크게 세 단계로 나누어진다. 주어진 객체의 반경 ε내에 이웃한 객체를 찾는다. 데이터베이스로부터 core point 조건을 만족하는 임의의 오브젝트을 선택하여 Seed로 잡는다. Seed로부터 밀도-도달가능한 모든 오브젝트들을 검색하여 cluster로 포함시킨다.

본 논문에서 사용하는 DBSCAN방법을 수행하기 위해서 적절한 파라메터가 요구된다. 사용되는 파라메터는 공간 객체들의 집합인 SetOfObjects, 그리고 클러스터에 포함되는지 여부를 판단하는 이진 프레디킷인 NP(Neighborhood Predicate)과 초기 클러스터를 생성하기 위한 MinNP가 사용된다. 이진 프레디킷인 NP는 이웃한 공간 객체가 존재하는 지를 판별하는 함수이다.

하나의 공간 객체 p로부터 r만큼의 거리 내에 공간 객체(점 또는 다각형)들이 포함되는지를 판별하여 발견된 객체(Object)의 개수가 MinNP를 만족하면 이웃한 집합인Seed로 결정한다. 오브젝트 p에 이웃한 오브젝트들에 대해 앞의 작업을 반복적으로 수행하여 만족하는 오브젝트의 집합을 클러스터로 만들게 된다. 밀도 기반 클러스터링방법은 〈알고리즘 2〉와 같다. 〔적용 예제1〕을 통하여 제안하는 공간 특성화의 수행방법에 대하여 알아본다.

알고리즘 2. 밀도 기반 알고리즘
Algorithm. 2 Density-based clustering

Precondition
: All objects in D are unclassified.
Input Parameter
:SetOFObjects, NP, MinNP

```
DBSCAN(SetOFObjects, NP, MinNP)
  ClusterID := nextID(NOISE);
  FOR I FROM 1 TO SetOfObjects.size DO
    Objects := SetOfObjects.get(i);
    IF Objects.CID = UNCLASSIFIED THEN
      IF Expandcluster(SetOfObjects, Objects, ClusterId
                       , NP, MinNP) THEN
        ClusterId := nextId(ClusterId)
      END IF
    END IF
  END FOR
END; // DBSCAN
```

〔적용 예제 1〕 인천 지역에 대한 여성 거주자를 대상으로 소득, 학력과 나이에 대한 특성화를 수행하시오. SMQL 기반의 질의문은 아래와 같다.

```
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
```

위의 질의문에 대한 밀도기반의 클러스터링을 이용한 공간 특성화의 수행방법은 크게 4단계로 구성되어 동작한다.

〔단계1〕 사용자에 의해 주어진 위와 같은 질의문(SMQL)은 질의처리기의 파싱 과정을 통하여 토큰의 형태로 분할하고, 분할된 토큰을 각 모듈과 연결하여 작업 관련 공간 데이터 수집 과정을 수행한다. 공간 데이터베이스로부터 census 테이블의 annual_income, education, age 그리고 공간 오브젝트들이 튜플 단위로 수집된다. 여기에서 각 튜플은 하나의 공간 객체를 의미한다. 〈표 1〉은 공간 데이터베이스에 저장되어 있는 데이터를 보여주고 〈그림 4〉은

공간 특성화를 위한 사전데이터인 각각의 작업관련 데이터로 각 오브젝트가 내포하고 있는 년 소득, 나이, 학력의 비공간 속성을 보여준다.

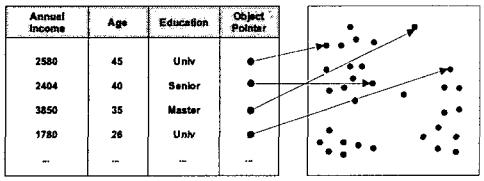


그림 4. 공간 및 비공간 데이터
Fig. 4 Spatial and non-spatial data

표 1. 작업관련 데이터
Table. 1 A task relevant data

id	annual_income	age	education	object	...
1	2580	45	Univ	⟨x, y⟩	...
2	2400	40	Senior	⟨x, y⟩	...
3	3850	36	Master	⟨x, y⟩	...
4	1780	26	Univ	⟨x, y⟩	...
5	3400	31	Univ	⟨x, y⟩	...
6	2300	22	Senior	⟨x, y⟩	...
7	...	...	...	...	...

〔단계 2〕 이 단계에서는 수집된 공간 및 비공간 속성들의 일반화 작업을 수행한다. 일반화 작업은 각 속성마다 구성된 개념 계층을 이용하여, 기존의 속성값을 상위 개념의 속성값에 대한 대치 작업인 데이터일반화 과정을 통하여 수행된다. 이러한 대치작업은 공통된 속성값들을 가지는 작업관련 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.

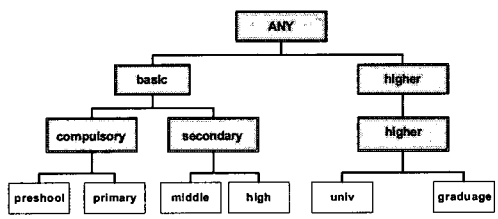


그림 5. 학력에 대한 개념 계층
Fig. 5 An education hierarchy

⟨그림 5⟩는 상위 개념의 속성값으로 대치과정을 위한 개념 계층을 보여준다. ⟨표 1⟩의 작업관련 데이터를 ⟨그림 5⟩의 개념 계층을 이용하여 ⟨그림 6⟩의 작업 관련 데이터의 일반화된 결과를 보여준다. 이러한 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. ⟨그림 6⟩의 객체 속성은 각 공간 오브젝트의 공간 정보를 나타내는 포인터(pointer)를 저장하며, 주어진 공간 영역에 대하여 다양한 지식발견을 위하여 공간 오브젝트들의 각 공간 속성을 이용하여 공간 군집화를 수행한다. 본 연구에서 사용한 군집화는 주어진 영역에서 밀도가 높은 지역에 대하여 군집화를 이루는 방법으로 대표적인 밀도 기반 클러스터링 방법인 DBSCAN을 기반으로 한다. ⟨그림 6⟩는 본 연구에서 제안한 밀도 기반 클러스터링의 군집화와 그와 관련된 일반화된 속성들에 대한 특성화의 결과를 보여준다.

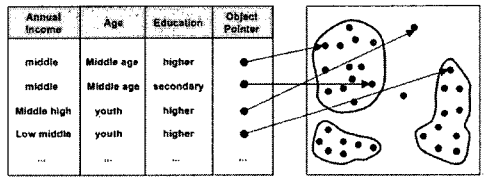
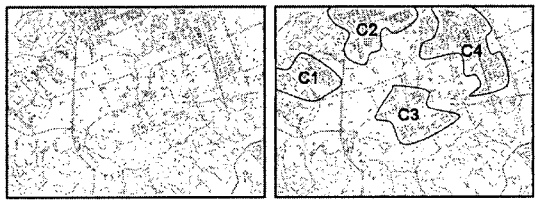


그림 6. 클러스터링을 적용한 특성화 결과
Fig. 6 A result of characterization based on Clustering



(a) (b)
그림 7. 예제 1에 대한 밀도 기반의 클러스터링
Fig. 7 Density-based clustering about example 1

주어진 공간 영역을 ⟨그림 7⟩의 (a)와 같이 인천지역의 모든 오브젝트들을 대상으로 특성화를 수행하는 것이 아니라, 밀도 기반의 클러스터링을 이용하여 공간 오브젝트가 밀집해 있는 영역을 군집화를 통해 탐사하고 발견된 영역에 공간 특성화기법을 사용한다. 즉, 사용자가 밀도가 높은 군집된 영역을 생성하기 위하여 클러스터 임계값을 기준으로 범위 안에 있는 공간 객체들을 각각의 클러스터로 만든다.

〈그림 7〉의 (b)와 같은 밀도 기반의 클러스터를 얻게 된다.

〔단계 4〕 본 단계에서는 위의 3단계의 결과를 이용하여 얻어진 튜플을 병합하는 과정을 거친다. 그리고 병합된 결과를 집계연산을 통하여 〔적용 예제 1〕에 대한 결과를 〈표 2〉와 같이 제공한다.

표 2. 공간 특성화 결과
Table. 2 A result of spatial characterization

cluster	annual_income	age	education	count
C 1	middle	middle age	Higher	481
C 3	middle	middle age	Secondary	316
C 1	middle high	middle age	Higher	156
C 2	low	youth	Higher	288
C 1	middle	youth	Higher	242
C 1	middle	teenage	Secondary	87
...	...	...	...	

IV. 공간 특성화 시스템의 구조 및 설계

본 논문에서 제안하는 밀도 기반 클러스터링을 적용한 공간 특성화 시스템은 다음 〈그림 8〉과 같다. 공간 특성화 시스템은 다음과 같은 구성요소를 가진다. 공간 특성화를 지원하는 데이터마이닝 질의 처리기〔10〕, 공간 특성화 엔진, 공간 데이터베이스, 공간 특성화를 지원하는 개념 계층 저장소이다. 이러한 공간 특성화를 위한 마이닝 시스템은 다음과 같은 특성을 지닌다.

첫째, 질의 처리기는 공간 특성화를 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.

4.1 공간 특성화를 위한 언어

사용자 인터페이스는 다음과 같은 기능을 수행한다. 사용자가 발견하고자하는 패턴에 대한 질의문을 유저 인터페이스를 통해 입력받는다. 작성된 질의문은 질의 처리기(Query Processor)의 파서를 통하여 처리한다. 공간 데이터마이닝을 위하여 사용자가 사용을 하는 질의언어는 본 연구의 사전 연구결과인 SMQL(Spatial Mining Query Language)을 사용하여 질의문을 작성한다〔10〕. 특성화를 위한 SMQL질의어의 BNF는 〈그림 9〉와 같다.

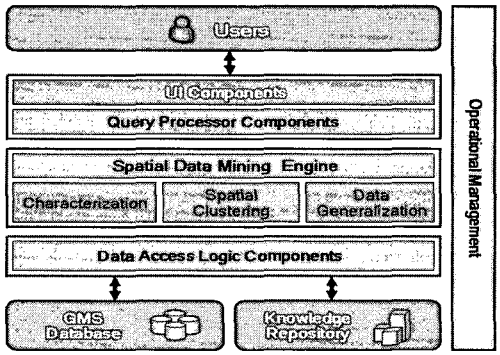


그림 8. 공간 특성화 시스템
Fig.8 Spatial Characterization System

```
SMQL_Query ::= MINE rule_header
[ USING HIERARCHY hierarchy_description ]
[ USING Clustering Distance conditions ]
[ FOR analysis_standards ]
FROM table_list
[ WHERE conditions ]
[ SET threshold_specification ]
```

그림 9. 공간 특성화 질의어의 BNF
Fig. 9 BNF of spatial characterization query

4.2 공간 특성화 엔진

제안된 공간 특성화 엔진은 지식을 탐색할 수 있는 방법으로 크게 공간 특성화와 밀도 기반 클러스터링 모듈로 구성된다. 두 가지 모듈 중 첫 번째로 공간 특성화 모듈은 공간 데이터베이스로부터 입력된 데이터들의 튜플들을 일반화하며, 밀도 기반 클러스터링 모듈은 공간적 위치속성을 가지고 있는 공간 객체를 대상으로 밀도 기반의 클러스터링을 통하여 공간분석을 위한 사전단계로 주어진 공간 영역을 각각의 클래스영역으로 나누어주는 모듈이다.

4.3 공간 데이터베이스

본 시스템은 GMS공간 데이터베이스를 이용한다〔9〕. 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 지원한다.

GMS는 데이터베이스 기술, 클라이언트/서버 기술 및 지리정보 기술이 통합된 개방형 공간 데이터베이스 관리 시스템으로 NT와 UNIX 서버를 위한 플랫폼을 제공하며 동일 클라이언트로 접속 가능한 데이터베이스이다. 또한, 기존의 관계형 데이터베이스는 공간데이터와 속성데이터를 분리하여 저장하는 것에 비해 GMS는 공간 데이터와 속성 데이터를 물리적으로 함께 저장하므로 지도 검색 및 공간 질의 시 빠른 속도를 제공한다.

4.4 공간 특성화를 지원하는 개념 계층 저장소

지식 저장소는 특성화 수행 시에 관련되는 지식이나 개념 계층을 저장한다. 즉, 공간 특성화를 지원하는 지식 저장소는 속성의 변환작업(즉, 일반화 작업)에 필요한 대치 속성값을 저장한다. 따라서 대치 속성값들은 공간 특성화 엔진의 공간 특성화 과정 중 데이터 일반화의 모듈 수행 시에 적절히 이용된다. 이러한 개념 계층의 구성을 시스템 사용자나 지식 전문가에 의해 제공할 수 있도록 공간 마이닝을 위한 질의 처리기와 함께 운용된다. 본 시스템에서는 사용자가 직접 개념 계층을 생성하여 지식 저장소에 저장할 수 있다. 〈그림 10〉은 사용자에 의해 생성되는 개념계층의 BNF를 보여준다.

```
Hierarchy_Generation
::= GENERATE_HIERARCHY hierarchy_name
    CONSTRUCT construction_methods
```

그림 10. 개념 계층을 생성하는 SMQL의 BNF
Fig.10 BNF of generating hierarchy

V. 실험 및 결과

본 연구에서 제안하는 기법의 성능분석은 정보이론(Information Theory)에서 정보의 불확실성을 포착하기 위해 사용하는 엔트로피(Entropy)를 이용하였다. 엔트로피는 (식 1)과 같이 정의된다.

$$Entropy(s) \equiv \sum_{i=1}^{C} -p_i log_2 p_i \quad \cdots\cdots (1)$$

위의 (식1)에서 S는 선택된 데이터의 집합, pi는 S가 i 클래스에 속할 비율, C는 집합 S와 구분 가능한 집합의 수를 각각 의미한다. 따라서, 위의 엔트로피와 각 속성에 대한 가중치에 대한 (식2)를 이용하여 (식3)의 정보이득을 계산할 수 있다.

$$W(\text{weight}) = wni/T \quad \cdots\cdots (2)$$

위의 (식2)에서 ni는 임의의 선택된 속성,wni는 ni가 가지는 가중치,T는 전체 가중치를 각각 의미한다. 공간 특성화의 결과가 가지는 정보이득 즉, 데이터의 분산지수인 Gain은 (식 3)과 같이 정의할 수 있다.

$$Gain(G) = E - W_a E_a + W_b E_b \quad \cdots\cdots (3)$$

여기에서 E는 전체 엔트로피, a는 공간 특성화에 사용되는 데이터, b는 전체 데이터 중 a를 제외한 나머지 데이터,W와 E는 각각 가중치와 엔트로피를 의미한다. (식3)의 Gain을 통하여, 선택된 집합 S의 모든 속성에 대한 가중치의 곱의 결과를 이용하여, 전체 속성에 대한 선택된 집합의 속성의 정보 이득을 얻을 수 있다.

〈알고리즘 3〉은 본 논문에서 제안한 밀도 기반 클러스터링 기법을 적용하여 얻어진, 각 공간 영역에 대한 특성화 결과의 정보이득을 구하는 알고리즘이다. 〈알고리즘 3〉의 총 수행시간은 O(n)이고, 데이터베이스의 ONE-PASS searching을 수행한다.

알고리즘 3. 정보이득 알고리즘
Algorithm. 3 Information Gain Algorithm

Precondition
: Value of Weight & Entropy
Input Parameter
: Attribute of Spatial Characteristic result

```
computeGain() //Count weight
FOR C FROM 1 TO NumberOfClass DO
    weight := getWeight(C)
    total weight T =: sum of weight
END FOR //Compute Entropy and Compute GAIN
WHILE
    compute Entropy(a)
    GAIN := Total Entropy - Entropy
END WHILE
RETURN GAIN
```

위의 〈알고리즘 3〉을 〔적용 예제 1〕에 적용한 결과를 〈표 3〉에서 보여주고 있다. 〈표 3〉의 (a)와 (b)는 각각 본 본문에서 제안한 공간 특성화방법과 GeoMiner를 이용하여 얻어진 정보이득의 결과를 보여주고 있다.

표 3. 특성화결과에 대한 정보이득의 비교
Fig.3 Comparing characterization result about Information result
(a) 제안하는 시스템 결과의 정보이득
(a) Information gain of proposed system

annual_income	age	education
0.841	1.269	0.986

(b) GeoMiner system 결과의 정보이득
(b) Information gain of GeoMiner system

annual_income	age	education
0.638	0.583	0.245

위의 〈표 3〉에서 (a)의 경우가 (b)의 경우보다 클러스터의 모든 공간 객체들의 속성들에 대해서, 높은 Gain을 가지며, 이는 본 논문에서 제안한 밀도 기반 특성화 방법이 기존의 방법보다 더욱 대표적인 정보를 특성화할 수 있었다. 또한, 동일한 공간 영역을 대상으로 동일한 입력 파라미터를 사용하여 기존 시스템과 제안하는 공간 특성화를 수행시킨 결과, 본 논문에서 제안한 밀도 기반의 클러스터링을 이용한 공간 특성화 시스템이 다양한 질의에도 높은 정보이득의 결과를 보여주었다.

VI. 결론

최근 지리정보시스템(GIS)분야에서는 방대한 양의 데이터들을 가지는 공간 데이터베이스로부터 규칙적이고 대표적인 특성 혹은 패턴들을 찾아내는 연구를 활발히 진행하고 있다. 그러나 기존의 공간 특성화방법들은 사용자가 미리 선택한 제한된 영역에 대한 특성화만을 수행하며, 공간적 밀집도가 적은 경우에도 공간적 밀집도를 고려하지 않고 공간 특성화를 수행하여, 특성화 지식의 효율성을 저하시키는 문제점을 가진다.

따라서, 본 연구는 밀도 기반 클러스터링을 적용하여, 특성화 작업을 수행하고자 하였다. 제안된 방법은 공간적 밀도를 고려하여 군집화하고, 이를 기반으로 비공간 속성에 대해서 특성화작업을 수행한다. 또한, 제안된 방법을 이용하여, 사용자에게 높은 유용성을 가지는 특성화 지식을 탐사할 수 있는 공간 특성화 시스템을 설계 및 구현을 하였다.

향후, 본 연구의 방법을 기반으로 공간상 존재하는 공간 객체간의 거리를 장애물(예를 들어, 산, 강, 도로 등)을 고려한 거리계산 방법을 이용하여, 특성화 결과의 유용성을 더욱 높이고자 한다.

참고문헌

〔1〕 M. Ester, H. -P. Kriegel, and B. Seeger. "Knowledge discovery in large spatial databases : Focusing Techniques for Efficient Class Identification." In proc. 4th Int'l. Symp. on Large Spatial Databases(SSD'95), pp.67-82, Portland, Maine, August 1995.

[2] 조성훈, 안동규, 김재홍, "CEO의 효율적/유효적 의사 결정을 위한 경영성과 데이터마이닝 시스템의 구축. 한국컴퓨터정보학회 논문지 5권 4호", pp 41-47, 2000.
[3] 오석, "지리 정보 추출의 자동화 알고리즘. 한국컴퓨터 정보학회 논문지 5권 4호", pp 21-27, 2000.
[4] M. Ester, H. -P. Kriegel and J. Sander "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, pp.160-187, 2001
[5] J. Han, and Y. Cai, and N. Cercone, "Knowledge Discovery in Databases : An Attribute-Oriented Approach." Proceedings of the 18th VLDB Conference. Vancouver, British Columbia, Canada, pp.547-559, 1992
[6] 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), pp.553-556, 1997
[7] M. Ester, H. -P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", In Proc. of ACM SIGMOD 3rd International Conference on Knowledge Discovery and Data Mining, pp.226-231, AAAI Press, 1996
[8] W. Lu, J. Han, and B.C.Ooi, " Discovery of General Knowledge in Large Spatial Databases", In Proc. Far East Workshop on Geographic Information Systems, Singgapore, June 1993.
[9] 박상근, 박순영, 정원일, 김명근, 배해영, "GMS : 공간 데이터베이스 관리 시스템", 공동 춘계학술대회, pp.217-224, 2003
[10] 박선, 박상호, 안찬민, 이윤석, 이주홍, "SIMS를 위한 공간 데이터 마이닝 질의 언어", 한국정보과학회 춘계 학술발표논문집 제 31 권 제 1 호, pp.70-72, 2003

저 자 소 개

유 재 현

2004년 2월 : 단국대학교 전자계산학과 학사
2004년 ~ 현재 : 인하대학교 대학원 석사과정
관심분야 : 데이터마이닝, 데이터베이스

이 주 홍

2001년 2월 : 한국 과학 기술원 컴퓨터 공학 박사
2002년 ~ 현재 : 인하대학교 컴퓨터공학부 부교수
관심분야 : 데이터마이닝, 데이터베이스, 정보검색, 신경망, 기계학습

박 태 수

2004년 2월 : 국립 공주대학교 정보통신공학과 학사
2004년 ~ 현재 : 인하대학교 대학원 석사과정
관심분야 : 데이터마이닝, 데이터웨어하우스

안 찬 민

2003년 2월 : 인하대학교 컴퓨터공학과 학사
2003년 ~ 현재 : 인하대학교 대학원 박사과정
관심분야 : 데이터마이닝, 알고리즘, 프로그램언어 및 설계

박 상 호

2002년 2월 : 인하대학교 컴퓨터공학과 학사
2004년 2월 : 인하대학교 대학원 컴퓨터정보공학과 석사
2004년 ~ 현재 : 인하대학교 대학원 박사과정
관심분야 : 데이터마이닝, 데이터웨어하우스

홍 준 식

2002년 2월 : 충북대학교 전기공학과 공학 박사
2004년 ~ 현재 : 영동대학교 전자의용공학부 전임강사
관심분야 : 신호처리(영상 및 음성), 패턴인식, 신경망