공간 데이터 마이닝 시스템의 설계 및 구현

(Design and Implementation of a Spatial Data Mining System)

배 덕 호* 백 지 행* 오 현 교* 송 주 원*
(Duck-Ho Bae) (Ji-Haeng Baek) (Hyun-Kyo Oh) (Ju-Won Song)
김 상 욱** 최 명 회*** 조 현 주***
(Sang-Wook Kim) (Myounghoi Choi) (Hyeonju Jo)

요 약 GIS 기술의 발달로 많은 양의 공간 데이터가 축적됨에 따라 공간 데이터 마이닝의 중요성이 커지고 있다. 본 논문에서는 새로운 공간 데이터 마이닝 시스템 SD-Miner를 제안한다. SD-Miner는 크게 입력과 출력을 담당하는 사용자 인터페이스, 공간 데이터 마이닝 기능을 처리하는 데이터 마이닝 모듈, DBMS를 이용하여 데이터를 저장하고 관리하는 데이터 저장 모듈의 세 부분으로 구성된다. 특히, 데이터 마이닝 함수 모듈에서는 공간 데이터 마이닝의 주요 기법인 공간 클러스터링, 공간 분류, 공간 특성화, 시공간 연관규칙 탐사 기능을 제공한다. SD-Miner는 다음과 같은 특징을 가진다. SD-Miner는 사용자로 하여금 공간 데이터 마이닝뿐만 아니라 비 공간 데이터에 대한 마이닝도 가능하게 하며, 각 마이닝 함수들을 라이브러리 형태로 제공하기 때문에 다른 시스템에서도 쉽게 사용 가능하다. 또한, 마이닝 매개 변수들을 테이블의 형태로 입력받기 때문에 시스템의 범용성이 높다. 개발된 SD-Miner의 실용성을 규명하기 위하여 실제 공간 데이터를 이용한 데이터 마이닝을 수행함으로써 여러 가지 의미있는 결과를 도출한다.

키워드 : 공간 데이터, 공간 데이터 마이닝, SD-Miner

Abstract Owing to the GIS technology, a vast volume of spatial data has been accumulated, thereby incurring the necessity of spatial data mining techniques. In this paper, we propose a new spatial data mining system named SD-Miner. SD-Miner consists of three parts: a graphical user interface for inputs and outputs, a data mining module that processes spatial mining functionalities, a data storage model that stores and manages spatial as well as non-spatial data by using a DBMS. In particular, the data mining module provides major data mining functionalities such as spatial clustering, spatial classification, spatial characterization, and spatio-temporal association rule mining. SD-Miner has own characteristics: (1) It supports users to perform non-spatial data mining functionalities as well as spatial data mining functionalities intuitively and effectively; (2) It provides users with spatial data mining functions as a form of libraries, thereby making applications conveniently use those functions. (3) It inputs parameters for mining as a form of database tables to increase flexibility. In order to verify the practicality of our SD-Miner developed, we present meaningful results obtained by performing spatial data mining with real-world spatial data.

Keywords : Spatial Data, Spatial Data Mining, SD-Miner

1. 서 론

정보 및 컴퓨터 기술의 발달로 인해 최근에는 다양한 분야에서 데이터가 쏟아져 나오고 있다. 정보 기술의 빠른 발전은 업무자동화를 촉진시켜 엄청난 양의 데이터를 수집 보관하는 것을 가능하게 하였고, 전자적으로 수집되는 데이터의 양은 매년 기하급수적으로 증가, 축적되고 있다. 이렇게 수집된 데이터를 정보의 형태로 가공하지 않는다

† 본 연구는 국토해양부 첨단도시기술개발사업의 연구비지원(07국토정보C05)과 2007년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단(KRF-2007-314-D00221)의 연구비 지원을 받았습니다.

* 한양대학교 전자컴퓨터통신공학과, smith@agape.hanyang.ac.kr, haeng0219@ns-corp.com, rapkyo@agape.hanyang.ac.kr, juwon@hanyang.ac.kr
** 한양대학교 전자컴퓨터통신공학과 교수, wook@hanyang.ac.kr(교신저자)
*** 네이버 시스템(주) 모바일 사업부, cmh775@neighbor21.co.kr, espion@neighbor21.co.kr

논문접수 : 2008.08.28 수정일 : 1차 2008.11.18 심사완료 : 2008.11.24

면 그 자체로는 아무 의미가 없다. 과거 데이터의 양이 방대하지 않을 때는 전문가들이 단순 통계 기법이나 질의를 통해 데이터를 분석하였다. 이러한 분석 방법은 데이터의 양이 증가할수록 효율성이 떨어지고, 분석을 통해 얻을 수 있는 정보에 한계를 가져왔다. 정보 기술을 이용하여 데이터를 여과, 분석하고 그 결과를 해석하는 자동화된 방안이 필요함에 따라, 대용량의 데이터로부터 의미 있는 데이터를 추출하여 숨겨진 규칙들을 발견함으로써 새로운 지식과 정보를 창출하는 데이터 마이닝(data mining) 기법[1]이 필요하게 되었다.

지리학에서도 컴퓨터 기술의 접목과 원격탐사(remote sensing), 모니터링 시스템(monitoring system), 지리 정보 시스템(GIS: geographic information system), 범지구 위치결정 시스템(GPS: global positioning system) 등의 발달로 많은 양의 공간 데이터가 축적되고 있다. 또한 이 대용량 공간 데이터 내에 잠재해 있는 유용한 정보와 지식들을 추출하기 위한 연구들이 시도되고 있다.

공간 데이터는 기존 일반 데이터와는 다른 특징을 가지므로 일반적인 데이터 마이닝 기법을 공간 데이터에 적용하는 것에는 한계가 있다[2][3][4]. 먼저, 공간 데이터는 거리 정보 및 위상 정보를 가지며, 외형적인 구조도 데이터 별로 상이하다. 또한, 공간 객체들은 공간에 무작위적으로 분포하지 않고 서로 영향을 주며, 그 영향은 객체 간의 거리나 인접성이 높을수록 더 커진다. 즉, 객체들은 서로 연관성을 가지며 근접한 객체일수록 멀리 있는 객체보다 더 큰 연관성을 갖게 된다. 따라서 기존 연구들은 공간 데이터의 특성을 고려한 공간 데이터 마이닝 기법들을 제안하고 있다[3][5][6][7][8][9][10][11][12].

공간 데이터 마이닝을 위한 기법들은 많이 연구되고 있지만, 공간 데이터 마이닝을 위한 상용화된 툴의 개발은 미흡한 상태이다. 현재, 상용화된 데이터 마이닝을 위한 툴은 SAS의 Enterprise Miner[13], SPSS의 Clementine[14], IBM의 Intelligent Miner[15], Rapid-i의 Rapid Miner[16] 등과 같이 여러 가지 기법들을 지원하여 데이터 마이닝의 다양한 작업을 가능하게 하는 범용제품군과 Rulequest사의 C5.0[17]이나 Neuro Dimension사의 NeuroSolutions[18] 등과 같이 특별히 하나의 기법만을 지원하는 전용제품군까지 다양하다. 또한, 실험용 구현 툴로서 WEKA[19]가 존재한다.

그러나 위에서 언급한 데이터 마이닝 툴들은 일반적인 데이터 마이닝 기법만 지원할 뿐, 공간 데이터 마이닝 기법은 지원하지 않는다. 공간 데이터 마이닝을 위한 툴로써 GeoMiner[6]나 FlowMiner[20]가 제안되었지만, GeoMiner는 현재 상용화되지 않았으며 입력되는 데이터는 특정 형식을 따라야만 한다[6]. 또한, FlowMiner는 여러 가지 기법들을 지원하는 않고, 특정 문제를 해결하기 위한 하나의 기법만을 지원한다[20]. 공간 데이터의 특징이 고려된 공간 데이터 마이닝을 위한 상용화된 툴은 거의 존재하지 않는다고 할 수 있다. 그러므로 일반 데이터 마이닝보다 더 전문적인 지식이 요구되는 공간 데이터 마이닝을 실제 응용에서 더 쉽게 사용하고 적용하기 위해서는 공간 데이터 마이닝을 위한 툴 개발이 필요하다.

본 논문에서는 공간 데이터 마이닝의 네 가지 주요 기법인 공간 클러스터링(spatial clustering)[7], 공간 분류(spatial classification)[21], 공간 특성화(spatial characterization)[22], 시공간 연관규칙 탐사(spatio-temporal association rule)[23]의 알고리즘의 기능을 제공하는 툴인 SD-Miner를 설계하고 개발한다. SD-Miner는 제안된 공간 데이터 마이닝 기법들의 특징을 분석하여 그 기법들의 성능을 개선하고 효과적인 공간 데이터 마이닝을 수행할 수 있도록 한다. 또한 SD-Miner를 이용한 실제 마이닝 사례를 제시한다.

SD-Miner는 공간 데이터뿐만 아니라 비공간 데이터도 마이닝이 가능하며, 각 마이닝 함수를 라이브러리 형태로 제공하기 때문에 다른 시스템에서도 쉽게 사용이 가능하다. 또한 사용자의 입력을 테이블 형태로 입력받아 처리하기 때문에 시스템의 범용성이 높으며, 사용자의 의견을 반영한 마이닝을 할 수 있도록 설계되었다.

본 논문의 구성은 다음과 같다. 제 2장에서는 공간 데이터 마이닝의 개념에 대해 살펴보고, 공간 데이터 마이닝에 사용된 네 종류의 기법에 대해 설명한다. 제 3장에서는 본 논문에서 개발한 SD-Miner의 구조에 대해 설명한다. 제 4장에서는 공간 데이터 마이닝의 성능 향상을 위해 개선된 사항에 대해 논한다. 제 5장에서는 구현된 SD-Miner를 통해 실제 데이터를 마이닝한 사례를 제시한다. 마지막으로, 제 6장에서는 본 논문을 요약하고 결론을 내린다.

2. 공간 데이터 마이닝

본 장에서는 공간 데이터 마이닝의 개념에 대해 살펴본다. 그리고 SD-Miner에서 적용한 공간 데이터 마이닝 기법들의 개념과 특징에 대해 설명한다.

공간 데이터 마이닝이란 데이터 마이닝의 공간적 확장으로 해석할 수 있으며 공간 데이터가 가지는 공간적 특수성을 고려한 마이닝이라 할 수 있다. 즉, 공간 데이터 마이닝은 공간 데이터베이스 내에 잠재되어 있고 흥미로운 정보와 공간적 상관관계, 다양한 공간적 패턴을 찾아내는 과정이다[1]. 공간 데이터는 텍스트로 이루어진 일반 속성 정보 뿐 아니라 2, 3차원 공간에서 존재하는 점, 선, 면의 다양한 객체로 이루어진 공간 정보를 포함한다[9]. 공간 정보에는 위상 정보, 거리 정보가 함께 존재하며, 공간 객체의 형상까지 포함한다.

공간 데이터 마이닝에서는 일반적인 비공간 데이터 마이닝에서의 비공간 속성의 패턴뿐 아니라, 객체의 공간적 속성, 시공간적 관계에 대한 패턴이나 규칙성을 추출하기 위한 기법들이 필요하다. 본 장에서는 유용한 공간 정보를 추출하는데 가장 널리 사용되는 기본적인 4가지 마이닝 기법인 공간 클러스터링, 공간 분류, 공간 특성화, 시-공

간 연관규칙 탐사 방법에 대해 설명한다.

2.1 공간 클러스터링

공간 클러스터링은 지리적, 위치적 특성에 따라 관련성이 높은 데이터들을 같은 그룹으로 분류하는 기법이다[7]. 본 논문에서는 클러스터링 방법 중 밀도기반 기법의 하나인 DBSCAN[24] 방식을 공간 데이터에 대하여 동작하도록 확장한 GDBSCAN[7] 방식을 채택한다.

GDBSCAN 방식은 공간 데이터의 객체 사이의 거리를 공간 데이터 거리 측정함수를 사용하여 두 객체 간의 가장 가까운 지점을 선택 후 계산한다. GDBSCAN 기법은 사용자가 입력한 두 객체 간의 거리가 입실론거리 이내이면 eps-이웃으로 여기고, eps-이웃의 개수가 사용자가 입력한 일정 임계값(threshold) 이상일 때 그 기준이 되는 객체를 코어 객체라 하고, 코어 객체를 기준으로 eps-이웃들을 같은 클러스터로 묶는다. 묶인 클러스터는 eps-이웃들이 코어 객체인지의 여부에 따라 확장된다. GDBSCAN과 DBSCAN은 클러스터링을 수행하는 객체의 종류가 공간 속성을 가지는 객체라는 것일 뿐, 클러스터 구성 방법상에서 차이를 가지지 않는다.

GDBSCAN 방식은 다음과 같은 장점들을 갖는다. 첫째, 다른 클러스터링 방법과 달리 크기와 모양을 가지는 공간 객체들을 클러스터링 하기에 적합하며, 밀도기반 클러스터링 기법을 이용하기 때문에 사람이 판단한 경우와 가장 유사한 클러스터를 구성한다. 따라서 공간 데이터의 특성을 가장 잘 반영하여 클러스터링을 수행할 수 있다. 둘째, GDBSCAN방식은 클러스터를 구성할 때 잡음의 영향을 거의 받지 않는다. 셋째, GDBSCAN은 원 스캔 방식으로 클러스터링을 수행하기 때문에 대용량의 공간 데이터베이스를 처리하는데 효과적이다.

2.1 공간 분류

공간 분류 기법에서는 객체의 일반 속성 뿐 아니라 객체들의 공간적 관계를 정의하기 위한 공간 속성도 고려하여 객체를 분류한다[21]. 공간 분류 기법은 일반 분류 기법과 동일하게 의사결정트리를 이용한다. 공간 분류와 일반 분류의 가장 큰 차이점은 공간 분류는 공간 객체의 인접지역의 집계 값(aggregation value)까지 이용한다는 것이다.

본 논문에서 공간 분류 기법은 [21]에서 제안된 공간 분류 기법을 사용한다. 이 기법[3][21]은 기존에 이웃 관계 그래프만을 이용했던 방식에서, 공간 객체들 간의 관계에 관한 술어도 검사 속성으로 사용하여 의사결정트리를 구축하는데 이용한다. 여기에서는 이단계의 분류를 위한 과정을 거친다. 일단계로 공간 분류를 위한 공간 속성을 객체들 간의 관계에 관한 술어로 표현한 후 이를 RELIEF 알고리즘[21]을 통해 이 술어 중에서 분류의 기준으로 삼을 수 있는 술어를 추출하고, 이단계로 추출된 술어들로 의사결정트리를 구축한다.

SD-Miner에서 사용하는 공간 분류의 장점은 다음과 같다. RELIEF 알고리즘을 통해 비효과적인 속성들은 미리 삭제되고 분류에 영향을 크게 미치는 속성들로 이진의 사결정트리를 구성함으로써 의사결정트리 구성에 소요되는 비용을 줄일 수 있다. 이진의사결정트리의 구축으로 이해하기 쉬운 규칙을 형성하고 RELIEF 알고리즘을 통해 가지치기의 컴퓨팅 비용을 줄임으로써, 더 빠르고 정확한 분류를 수행할 수 있다. 이러한 공간 분류 기법은 고객을 성향이나 등급을 감안해서 지역적으로 분류한 후 차별화된 고객을 유치하고자 할 경우에 적용될 수 있다.

2.3 공간 특성화

공간 특성화 기법은 공간상에 주어진 공간 객체와 각 객체가 내포하고 있는 데이터 집합으로부터 탐색하고자하는 공간 영역에 대한 데이터 클래스의 전체적인 윤곽을 파악하는 방법으로, 사용자에게 간략하고 간결한 요약정보를 제공한다[22]. SD-Miner에서는 [22]에서 제안된 공간 특성화 기법을 사용한다. 이를 위해 사용자가 지식이나 패턴을 발견하고자 하는 영역 범위에 대한 공간 및 비공간 데이터를 공간 데이터베이스로부터 먼저 수집한다.

공간 특성화 기법에서는 주어진 공간 객체의 특성이 이웃된 공간으로 확장되는지 판단한다. 이때, 공간 특성화의 대상이 되는 공간 객체들은 거리 또는 방향을 통해 서로의 이웃으로 정의된다. 이웃으로 정의된 객체들은 이웃 정보 테이블에 저장되어 관리된다. 공간 특성화의 대상이 되는 공간은 공간 확장 알고리즘에 따라 확장될 수 있다. 이를 위하여 미리 생성된 이웃 정보 테이블을 이용한다.

공간 특성화 기법의 장점은 분석된 특성화 패턴을 공간적으로 확장 적용한다는 것이다. 타겟 객체의 특성을 공간적으로 확장하면서 인접한 주위의 특성을 함께 파악할 수 있으며, 반대로, 확장에 따른 특성의 변화를 파악할 수 있다. 이렇게 함으로써, 타겟 객체와 관련성이 높은 지역의 특성까지 함께 파악 분석하고, 변화되는 패턴들을 찾을 수 있다.

2.4 시공간 연관규칙 탐사

공간 연관규칙 탐사 기법을 사용하면 공간 데이터와 공간 데이터, 공간 데이터와 비공간 데이터 사이의 특정한 상관관계를 분석하여 공간적 객체들 사이의 위상적 상관관계와 공간 객체들 사이의 거리 관계를 표현할 수 있다. 여기에 시간 데이터의 분석을 통해 시간에 따른 시공간 연관규칙 탐사도 할 수 있다. SD-Miner에서는 [23]에서 제안된 공간 연관규칙 탐사 기법을 사용하며, 추가로 이 기법을 확장 제안한 시공간 연관규칙 탐사 기법을 사용한다.

공간 연관규칙 탐사를 위해서는 공간 객체들 간의 공간 관계가 정의되어야 한다. 공간 객체들의 공간 관계는 공간적 술어로 표현되며, 공간적 술어는 제 3장에서 설명하는 SD-Miner의 데이터베이스 안에 개념 계층 데이터로 사

용자가 미리 정의하여 저장된다. 공간 객체들 간의 관계가 공간 술어로 정의되어 있다면, 그 정의된 술어는 비공간 속성과 같이 하나의 속성으로 보고, Apriori 알고리즘을 사용하여 시공간 연관규칙을 탐사한다.

공간 연관규칙 탐사 기법의 장점은 공간적 술어이냐 비공간 술어이냐에 제약을 받지 않고 여러 가지 규칙을 찾을 수 있다는 것이다. 즉, 공간 술어와 공간 술어의 연관규칙 탐사뿐만 아니라 공간 술어에서 비공간 술어의 연관규칙 탐사를 할 수도 있다. 또한, 비공간 술어에서 공간 술어의 연관관계를 찾을 수도 있다. 또 다른 이 기법의 장점은 공간 객체의 타입이 공간적 개념 계층으로 구성되어 있어, 수준별 다양한 연관규칙 탐사를 도출한다는 것이다. 예를 들어, '아파트가 강주위에 위치해 있으면 집값은 비싸다.' 라는 규칙으로 부터 '아파트가 한강 주위에 위치해 있으면 집값이 비싸다.'라는 더 세분화된 규칙을 얻을 수 있다. 이렇게 수준별 연관규칙 탐사를 통해 더 다양한 연관규칙 탐사를 할 수 있다.

3. SD-Miner 구조

본 장에서는 본 논문에서 설계하고 개발한 공간 데이터 마이닝 시스템 SD-Miner(Spatial Data Miner)의 구조에 대해 설명한다. 그림 1은 SD-Miner의 구조를 나타낸다. SD-Miner는 크게 그래픽 사용자 인터페이스(GUI: graphic user interface), SD-Miner 모듈, DBMS(database management system)관리 모듈의 3가지 부분으로 이루어진다.

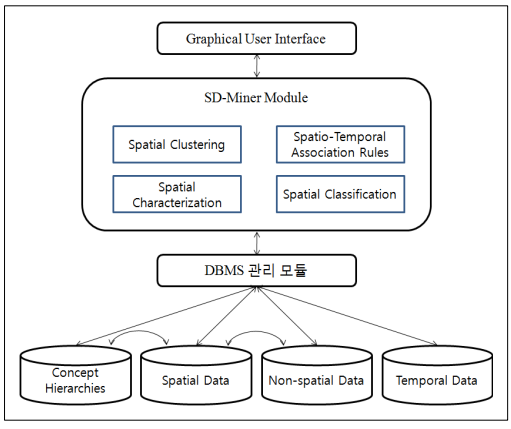

그림 1. SD-Miner 구조

첫째, GUI는 마이닝을 위한 여러 가지 매개변수들을 입력받아 SD-Miner 모듈에 전달하고 마이닝 된 결과를 테이블이나 차트, 지도 등의 형태로 보여주는 역할을 한다. SD-Miner에서 사용되는 각각의 함수는 함수 별로 입력 매개변수와 결과가 조금씩 차이가 있으며, 각각의 모듈별로 입력 및 결과형식이 다르게 나타난다.

둘째, SD-Miner의 핵심 모듈인 SD-Miner 모듈은 실제 마이닝 함수가 수행되는 모듈이며 GUI에서 입력받은 값들을 이용하여 마이닝된 결과를 DBMS에 전달하는 역할을 한다. 마이닝 함수로는 제 2장에서 설명한 4가지 함수가 제공된다.

셋째, DBMS 관리 모듈은 데이터가 저장된 데이터베이스를 관리하고, SD-Miner 모듈과 데이터베이스의 연동을 도와주는 역할을 한다. DBMS 관리 모듈은 오라클 10g[25]를 사용하며 여기서 제공되는 공간 함수를 사용하여 SD-Miner의 성능을 높여주고 알고리즘의 복잡도를 줄인다.

SD-Miner는 DBMS관리 모듈에 4가지 형식의 데이터를 저장하고 사용한다. 이들은 공간 정보를 포함하는 공간 데이터, 공간 정보외의 비공간 속성을 포함하는 비공간 데이터, 연관규칙 탐사에서 시간개념으로 사용될 시간 데이터, 공간적 술어가 저장된 개념 계층 데이터이다.
공간 정보를 포함하는 공간 데이터는 객체의 거리 정보와 위상정보를 가지고 있으며, 외형적인 구조가 함께 표현된다. 공간 데이터는 오라클 10g에서 제공하는 공간 데이터 저장 형식인 SDO_GEOMETRY 형식으로 저장된다.

객체의 비공간 속성을 저장하는 비공간 데이터는 저장된 공간 객체들의 비공간 속성이다. 예를 들어, 공간 객체가 주택의 타입을 갖는다면, 주택은 주택 구성원 수, 주택 소유자 학력, 주택 소유자 연봉, 보유 차량 수 등의 다양한 비공간 속성 값을 가질 수 있다.

객체의 시간 속성을 나타내는 시간 데이터는 시공간 연관규칙 탐사를 위해 사용되며, 비공간 속성과 같은 형태로 시간의 흐름에 따른 데이터 값을 갖는다. 시간 데이터의 예로는 주택의 월별 전력사용량, 계절별 전력사용량 등을 들 수 있다.

개념 계층 데이터는 공간 객체들 간의 관계를 정의해주는 공간적 술어 데이터이다[23]. 공간적 관계는 공간 객체들 사이의 거리에 기반한 위상적 관계를 표현한다. 예를 들어 하나의 공간적 관계는 그림 2와 같은 개념 계층을 가지며, 각각의 술어는 객체 사이의 거리로 정의된다. 개념 계층은 계층별로 상위 술어가 하위 술어를 포함하며 계층별로 술어와 술어가 정의되는 거리와 술어의 상위 개념이 테이블 형태로 저장된다.

개념 계층은 전문가에 의해서 주어지거나 사용자에 의해서 생성될 수 있으며, 공간 객체들의 이웃에 관한 공간적 관계는 이 개념 계층에서 제시된 거리에 기반하여 형성된다. 개념 계층에서는 하나의 술어 내에 여러 개의 세부적인 술어가 있을 수 있다. 예를 들어, '두 객체가 이웃한다'와 같은 공간적 술어는 '두 객체가 인접한다', '근처에 있다', '교차한다'와 같이 세부적인 공간 술어를 하위 개념으로 포함한다. 이러한 계층은 반드시 그림 2를 따라야 하는 것이 아니며, 사용자의 입력에 따라 다르게 정의될 수 있다.

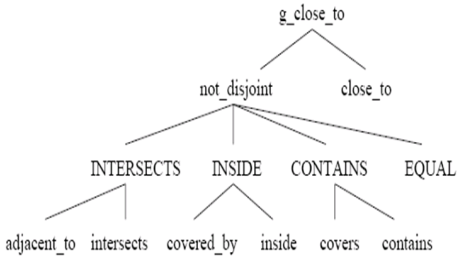

그림 2. 공간 술어 개념 계층[23]

공간 객체를 마이닝에 사용하기 위해서는 공간상의 거리나, 공간 객체의 면적 측정 등의 공간 측정 함수가 필요하다. SD-Miner에서는 공간 객체의 객체들 사이의 관계를 알아보기 위해 오라클 10g에서 제공하는 공간 함수를 사용한다[25]. 비공간 데이터와 시간 데이터를 마이닝에 사용하기 위해서는 데이터들을 마이닝에 이용 가능하도록 가공하는 작업이 필요하다[26]. 이러한 가공 작업을 위해서, SD-Miner는 마이닝에 사용할 속성과 그 속성의 가공된 데이터에 대한 정보를 모두 사용자로부터 테이블 형태로 입력받아 처리한다. 이렇게 함으로써, SD-Miner는 입력 데이터에 제한 받지 않고, 시스템의 범용성을 높이고 있다.

다른 마이닝 시스템과 구분되는 SD-Miner의 장점은 다음과 같다. 첫째, 각 공간 데이터 마이닝 함수는 일반 데이터 마이닝 함수를 확장한 것이며 공간 데이터 뿐 아니라, 비공간 데이터도 마이닝이 가능하다. 이때, 사용자의 특별한 입력사항 없이 데이터의 입력 형태에 따라 시스템에서 공간데이터인지 비공간 데이터인지를 판단하여 자동적으로 데이터 마이닝을 수행하게 된다.

둘째, 각 마이닝 함수를 라이브러리 형태로 제공한다. 따라서 SD-Miner에서 제공하는 마이닝 함수들은 타 시스템에서도 사용이 가능하다. 반대로, SD-Miner에 타 시스템에서 제공하는 마이닝 함수들의 추가도 용이하다.

셋째, 사용자의 입력을 테이블 형태로 입력받아 처리하기 때문에 시스템의 범용성이 높다. SD-Miner에서는 일부 인자 값을 제외하고 데이터베이스에 저장된 테이블의 이름을 입력받는다. 이때, 시스템 상에서는 자동으로 기본키와 기타 필요한 속성들을 검색하여 작업을 수행한다.

사용자의 입력을 테이블 형태로 입력받음으로써, 사용자는 SD-Miner에서 제공하는 GUI를 이용하지 않고, 직접 데이터 마이닝 모듈의 연동이 가능하다. 이는 SD-Miner에 다양한 GUI가 연동 가능함을 의미한다. 또한, 공간 술어의 개념 계층이나 각 데이터 마이닝 기법에 필요한 정보 등 복잡한 데이터를 테이블 형태로 입력받음으로써, 사용자가 매번 직접 입력해야 하는 번거로움을 줄일 수 있으며, 데이터 입력 시 사용할 응용 프로그램의 선택에 있어 자유도를 제공한다.

넷째, 공간 술어를 정의할 때 사용자의 의견을 반영한다. 일반적으로 공간 데이터는 각기 다른 축적 값과 데이터 분포도를 가지고 있다. 두 객체간의 거리차가 1인 경우 지도의 축적 값에 따라 거리 1은 1킬로미터가 될 수도 있고, 100킬로미터가 될 수도 있다. 이러한 상황에서 모두 같은 조건의 공간 술어를 사용한다면, 공간 술어의 한 예인 근접하다는 의미 자체가 달라질 수 있다. 즉, 데이터에 따라, 또는 사용자의 정의 기준에 따라 공간 술어는 자유롭게 변경 가능하여야 한다. SD-Miner에서는 공간 술어를 정의할 때 개념 계층의 테이블을 이용하여 원하는 수준의 공간 술어 정의를 사용자가 할 수 있도록 한다.

다섯째, 시스템 구현 시 오라클 10g에서 제공하는 함수를 이용함으로써, 구현이 단순화되고, 수행 속도가 빠르다. 오라클 10g에서 제공하는 공간 함수들을 SQL 쿼리를 사용하여 처리하며, 수행 속도 향상을 위한 별도의 인덱스 구현 과정이 생략되어 알고리즘 구현이 단순화되었다. 이러한 방법으로 알고리즘의 구현 복잡도를 줄이고, 보다 정확한 공간 데이터에 대한 여러 측정값을 얻을 수 있다.

4. SD-Miner의 설계 및 구현

본 장에서는 제 2장에서 설명된 기존의 공간 데이터 마이닝 함수의 문제점을 지적하고 이를 보완하기 위한 새로운 방안을 제안하며, 제안된 방안을 적용한 시스템의 설계에 대해 설명한다.

4.1 공간 클러스터링

제 2.1절에서 설명한 GDBSCAN 방식은 많은 장점을 갖고 있음에도 불구하고, 일정 거리이하에 존재하는 공간 객체의 개수만을 고려한다는 문제점이 있다. 즉 공간 객체의 개수만을 고려하기 때문에 크기나 모양을 가지고 있는 공간 데이터의 특성을 반영하지 못한다.

본 논문에서는 효과적이고 정확한 클러스터링을 위해 다음과 같은 방법들을 제안한다. 첫째, 코어 객체 판단 시, 개수뿐만 아니라 면적도 이용하여 코어 여부를 판단한다. GDBSCAN 방식에서는 공간 데이터가 위치 정보와 함께 데이터의 모양과 면적, 지형적 특성을 가진다는 특성을 반영하지 않고, 일정 거리 이하에 존재하는 객체 개수를 코어 객체의 판단기준으로 본다. 우리는 일정 거리 이하에 존재하는 객체들의 면적의 합이 전체 공간 면적의 일정 비율 이상이 되면 그 중심 객체를 코어 객체로 여긴다. 여기서 일정 비율이란 면적의 합이 사용자가 정의한 최소 면적 비율 이상을 만족하는 것을 말한다.

그림 3에서 세 개의 클러스터로 클러스터링이 수행된 것을 볼 수 있다. 이 때, 코어 객체를 판단하는 임계값의 개수가 3이라고 가정하면, 그림 3과 같은 클러스터가 구성이 되지만, 임계값의 개수가 4라고 가정하면 두 번째 클러스터인 Cluster2는 구성되지 못할 것이다. Cluster2와 같이 거리상 가까운 객체가 임계값의 개수보다는 작지만 그

객체의 면적이 다른 객체들의 면적보다 상대적으로 월등히 크다면 그 객체는 코어 객체가 되어야 한다. 따라서 클러스터 구성의 확장 시 코어 객체의 판단 기준은 개수와 면적 중 하나만 만족하는 경우도 고려하여야 한다. 실제 공간 데이터로 테스트 결과, 코어 객체 판단 기준을 후자로 했을 때의 클러스터링 결과가 사용자가 원하는 결과에 더 근접하게 나타나는 것을 볼 수 있었다.

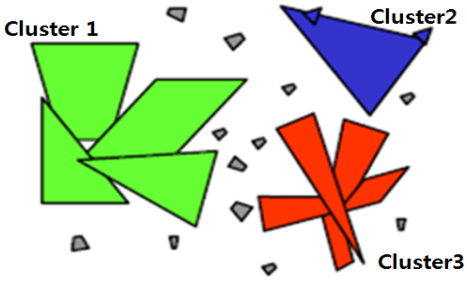

그림 3. 공간 객체 면적 이용 예

SD-Miner에서는 사용자의 판단에 따라, 개수만을 이용하여 클러스터링을 수행할 수도 있고, 면적만을 가지고 클러스터링을 수행할 수도 있으며, 이 두 가지 중 하나만 고려하여 클러스터링을 수행할 수도 있다. 즉, SD-Miner에서는 면적을 이용해야 한다는 강제성이 아닌 사용자의 선택에 의해 보다 나은 결과를 얻을 수 있도록 자율성을 부여했다.

둘째, 본 논문에서는 클러스터링의 효율성을 높이고, 한 번도 클러스터의 수행과정에 포함되지 않은 객체만을 대상으로 코어 여부를 판단하는 방법을 제안한다. 클러스터를 구성하기 위해 객체들을 판별 시, 모든 공간 데이터에 대해 매번 검색하게 되면, 그 오버헤드는 엄청나게 된다.

본 논문에서는 기본적으로 단 한 번의 스캔을 수행하면서, 먼저 클러스터 생성 과정에서 처리되었던 데이터들에 대해서는 다음번에 다시 고려되는 일이 없도록 한다. 이를 위해 먼저 하나의 테이블을 생성하여 이 테이블에 클러스터링 대상이 되는 객체들의 아이디를 저장한다. 저장된 객체 아이디는 각각 클러스터 아이디를 부여 받게 된다. 한 번도 클러스터링의 과정이 진행되지 않았을 때의 모든 객체의 클러스터 아이디는 널 값이다. 클러스터링을 위해 먼저 클러스터 아이디가 널인 객체 하나를 선택하여 그 객체의 코어 여부를 판단한다. 이 객체가 잡음 객체라면 이를 표시하며, 이 객체가 코어 객체라면 클러스터 아이디가 부여 된다. 이 때, 이웃되는 객체들의 클러스터 아이디도 모두 부여됨으로써, 다음 코어 객체 판단을 위한 객체의 후보 수가 줄어든다. 이렇게 함으로써, 클러스터 아이디가 이미 부여된 객체들은 코어의 후보에서 제외되며, 전체 수행 시간을 줄일 수 있게 된다.

예를 들어, 그림 4와 같이 1에서 10까지의 객체 아이디를 갖는 10개의 객체에 대해 클러스터링을 수행한다고 하자. 먼저, 클러스터 결과 테이블이 생성이 되고, 그 클러스터 결과 테이블에는 10개의 객체에 대한 객체 아이디가 부여되고, 클러스터 아이디는 널 값으로 저장된다. 코어 객체를 찾기 위하여 클러스터 아이디 중 널 값인 객체를 하나 선택한다. 1번 객체가 선택이 되어 코어 객체라 판별이 되고, 이때 같은 클러스터에 2, 4, 5, 7번의 객체가 포함되도록 선택이 되었다면, 1, 2, 4, 5, 7번 객체는 같은 클러스터 아이디를 부여 받게 된다. 다음으로, 선택된 각각의 2, 4, 5, 7번 객체에 대해 코어 여부를 판단하게 되고 더 이상 확장이 되지 않는다면, 클러스터 결과 테이블에서 다음 클러스터 아이디 값이 널인 3번 객체를 선택하게 된다. 3번 객체의 코어 객체 판별 후, 6, 8, 9, 10번 객체가 같은 클러스터에 포함된다면 공간 클러스터링 과정이 모두 끝나게 된다. GDBSCAN 방식을 보완한 이 방법은 클러스터링 수행속도를 향상시키며 효율성, 정확도를 높이게 한다.

SD-Miner에서는 클러스터를 구성할 수 있는 중심이 되는 코어 객체를 기준으로 클러스터를 확장하여 클러스

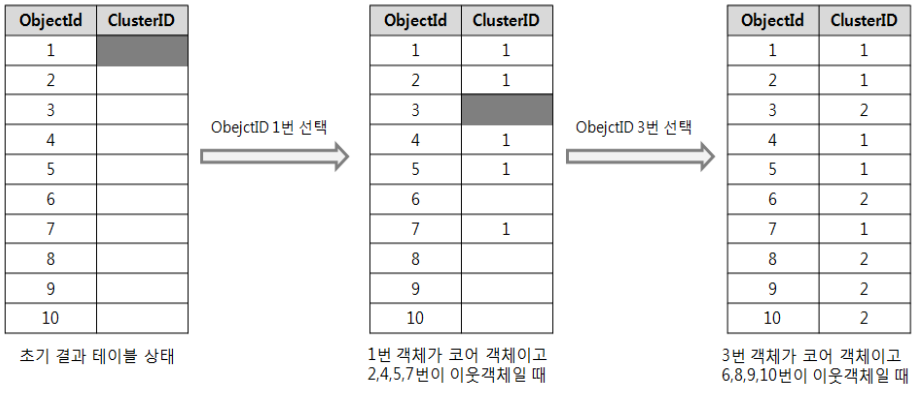

그림 4. 클러스터링 수행 예

터를 구성한다. GDBSCAN 방식을 보완한 공간 클러스터링 수행 과정은 그림 5와 같다. 먼저, 데이터베이스에 저장되어 있는 각각의 공간 객체, 비공간 객체, 매개변수들을 입력받는다. SD-Miner에서의 공간 클러스터링에서 필요한 매개변수는 같은 클러스터로 묶일 수 있는 객체간의 최소 거리인 입실론 값과, 코어 객체를 판단하기 위한 클러스터의 최소 개수, 최소 면적, 코어 여부를 판단하는 기준이다. 여기서, 코어 여부를 판단하는 기준은 면적, 개수, 면적 또는 개수 중 하나라도 만족하는 경우 세 방식 중에서 선택한다. 다음, 입력받은 객체들을 분석하여 클러스터링에 필요한 정보들을 추출한다. 그리고 추출된 정보들을 이용하여 결과가 저장될 테이블을 생성하고, 클러스터링을 수행한다.

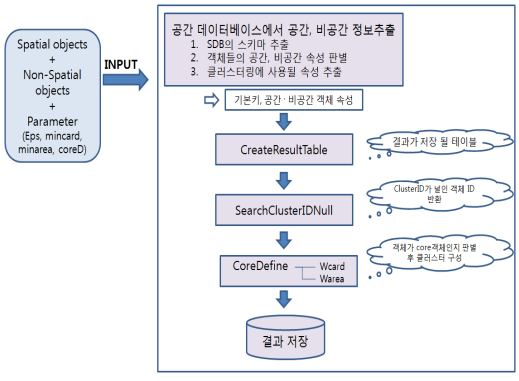

그림 5. 공간 클러스터링의 과정

4.2 공간 분류

제 2.2절에서 설명된바와 같이 본 논문에서는 의사결정트리를 구축하는데 이진의사결정트리를 사용한다. 이때, 트리 구축의 대상이 되는 객체들은 공간 술어에 대한 값을 가지고 있어야 한다. 각 객체들의 공간 술어들은 객체의 속성 값으로 저장되어 필요시 호출하여 사용되게 된다.

공간 분류를 위해서는 버퍼라는 개념을 사용한다. 버퍼란 한 객체를 중심으로 일정거리만큼 확장된 영역을 말한다. 공간 분류에 사용되는 공간 객체의 술어로 공간적으로 관련 있는 비공간 속성들을 사용하는데 예를 들면, 쇼핑센터라는 공간 객체의 술어로 다음과 같은 속성이 새롭게 만들어진다. '쇼핑센터를 중심으로 반경 1km 안의 인구 수.' 이러한 값을 집계 값이라고 하고, 이 값이 공간 속성의 술어로 선택된다. 이 때, 쇼핑센터가 버퍼가 되는 것이다. 버퍼를 사용하기 위해서는 버퍼 크기라는 변수가 필요하다. 버퍼 크기는 타겟이 되는 객체 주위의 확장되는 반경을 위한 것이다. 버퍼 크기의 값은 여러 가지 값들이 사용자에 의해 입력 된 후 정보 이득(information gain)이 가장 큰 술어가 있는 버퍼에 대한 크기를 선택한 후, 모든 객체에 같은 크기를 적용하게 된다. 이럴 경우 정보 이득이 가장 큰 술어 이외에 다른 술어들은 공간 분류에 대한 정확한 척도가 되지 않을 수도 있다.

따라서 SD-Miner에서는 사용자에게 먼저 버퍼 크기를 입력받고, 그 입력받은 크기를 통해 도출되는 모든 술어들에 대해서 정보 이득을 구하여, 각 버퍼 마다 정보 이득이 공간 분류를 위한 정확한 척도로 사용하도록 한다. 사용자는 다양한 버퍼 크기를 입력 한 후, 각 크기에 대해 모든 정보 이득을 구한 후 술어로 사용한다. 본 논문에서는 기존 방법과는 달리 사용자에게 버퍼 크기를 입력받아 도출되는 모든 집계 값을 술어로 사용하고, 모든 버퍼 크기에 대한 술어들에 대해서 정보 이득을 구한다. 그 값들이 공간 분류에 대한 척도로 사용가능 하도록 제안한다. 이에 의해, 버퍼 크기가 공간 객체마다 다른 값을 갖게 되지만, 서로 다른 집계 값을 갖는 술어들이 공간 분류의 척도로 사용될 수 있기 때문에 더욱 정확한 공간 분류를 할 수 있게 된다.

또한, SD-Miner에서는 다수결 투표(majority voting)의 개념을 사용함으로써, 이진의사결정트리의 크기를 결정한다. 공간 분류를 위한 이진의사결정트리의 구축 시, 트리의 방대한 확장을 막기 위해 사용자로부터 구성 트리의 크기를 입력받고, 입력받은 크기 이상으로 트리 구성이 진행되었다면 더 이상 트리의 확장을 하지 않는다. 예를 들어, 사용자가 80%라는 값을 입력하였을 때 전체 데이터 중 80%이상의 속성이 의사결정트리를 구성하는 경우 더 이상 트리 구축을 확장하지 않는다.

그림 6은 본 논문에서 제안하는 공간 분류의 과정을 보여주고 있다. 먼저 입력 받은 공간 객체와 트레이닝 셋 데이터를 이용하여 공간 분류가 진행이 된다. 그리고 공간 객체의 개념 계층 데이터를 이용하여 공간 객체 사이의 공간 술어를 정의한다. 정의된 수많은 공간 술어들 중에서 실제 분류에 이용이 되는 적절한 술어를 도출하기 위해 RELIEF 알고리즘을 수행한다. 이 알고리즘을 통해 얻어진 술어를 이용하여 이진의사결정트리를 구축한다. 이를 통해 우리는 공간 분류의 정확도를 향상시키고, 시간 복잡도를 줄일 수 있다.

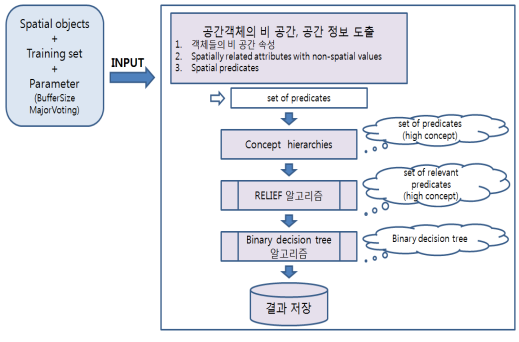

그림 6. 공간 분류의 과정

4.3 공간 특성화

제 2.3절에서 설명된 공간 특성화 방법에서 객체들의 이웃에 관한 정보를 갖고 있는 이웃 정보 테이블은 모든 공간 객체들을 대상으로 한다. 이는 O(n2)의 시간을 필요로 한다. 복잡도를 줄이기 위해서는 모든 객체들을 대상으로 이웃 정보 테이블을 구축하는 방법을 피해야 한다. 본 논문에서는 특정 공간 안에 있는 공간 객체들의 이웃들만을 대상으로 이웃 정보 테이블을 구축하는 방법을 제안한다. 이 방법은 정해진 공간별로 이웃 정보 테이블을 구축해야 하는 단점이 있지만, 공간 객체 개수가 많아지는 경우 모든 객체를 대상으로 이웃 정보 테이블을 만드는 것에 비해 처리 시간이 효율적이며, 만족스러운 처리 결과를 얻을 수 있다.

SD-Miner에서의 공간 특성화의 과정은 그림 7과 같다. 먼저 특정 공간에 대한 공간 객체의 공간 및 비공간 속성이 입력으로 주어져야 한다. 입력 데이터와 개념 계층 데이터를 이용하여 공간 객체들을 정의하고, 비공간 속성들의 빈번한 패턴을 찾는다. 객체의 공간 속성들은 공간 개념 계층의 술어에 의해 객체간의 관계로 표현된다. 객체의 패턴 분석을 위해 객체의 비공간 속성을 이용하여 속성 일반화과정을 거쳐 비공간 속성의 값을 계층 구조로 일반화하게 된다. 예를 들어, 학력 속성에 대해 초졸, 중졸, 고졸, 대학졸, 대학원졸 이라는 속성 값이 존재 할 때, 이를 초졸, 중졸, 고졸은 저학력으로 대학졸, 대학원졸은 고학력으로 일반화 시키면 더 일반화된 패턴을 찾을 수 있다. 그 다음, 타겟이 되는 객체들의 이웃 객체들에 대해 이웃 관계 테이블을 구축하고, 앞에서 찾아진 비공간 속성의 패턴들이 적용 가능한지 검사한다. 이웃 관계 테이블을 구축함으로써, 공간 확장의 범위를 조절할 수도 있다. 비공간 속성의 패턴들이 이웃 객체에 대해서도 발생한다면 또 다시 그 객체의 이웃들에 대해서도 반복하여 패턴들이 적용 가능한지에 대해 살펴본다. 이렇게 하여 최대 확장 가능 할 때까지 확장하여 적용하면 공간 특성화의 패턴을 찾을 수 있다.

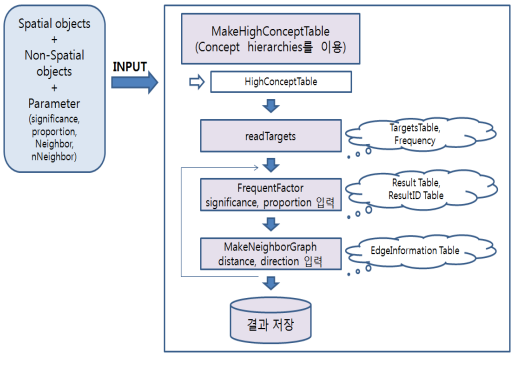

그림 7. 공간 특성화의 과정

4.4 시공간 연관규칙 탐사

제 2.4절에서 언급된 공간 연관규칙 탐사의 장점에도 불구하고, 공간 연관규칙 탐사적용을 위해서는 다수준 연관 규칙을 위한 적절한 임계값 조절과 공간 술어를 정의함에 있어 어려움이 있으며, 비공간 속성의 데이터 입력 형식에 따라 연관규칙 탐사에 차이가 나는 문제가 있다. 본 논문에서는 사용자에게 입력 데이터의 유연성을 제공하고, 보다 다양한 연관규칙 탐사의 분석을 위해 다음과 같은 방법들을 제안한다. 또한, 시공간 연관규칙 탐사를 위한 방안을 제시한다.

첫째, SD-Miner는 다수준 연관규칙 탐사를 위한 최소 지지도와 최소 신뢰도의 값을 자동으로 계산하기 위해 상위 수준의 사용자 입력 값을 이용하여 하위 수준의 임계값은 데이터의 분포 형태에 따라 적절한 수준으로 조절하는 방안을 사용한다. 공간 연관규칙 탐사를 위해서는 연관규칙 탐사의 타겟이 되는 가장 상위 단계의 공간 데이터 타입이 정의되어야 하며, 정의된 가장 상위 단계의 타겟에 대한 최소 지지도와 최소 신뢰도를 입력받아야 한다. 다음 타겟의 하위 단계로 내려갈수록 최소 지지도와 최소 신뢰도의 값도 낮아져야한다. 이는 더 포괄적인 의미의 상위 계층의 객체수의 합과 더 세분화된 하위 수준의 객체수가 다르기 때문이다. 따라서 하위 단계일수록 상위 계층의 객체 수에 비례하여 최소 지지도와 최소 신뢰도를 낮추어 주면 적절한 수준의 적용 가능한 임계값을 찾을 수 있다.

둘째, 비공간 속성의 데이터 값에 대한 일반화를 수행하여 연관 규칙을 분석하는 방안을 제안한다. 비공간 속성의 데이터는 데이터의 종류에 따라 다양한 값이 저장이 된다. Apriori 알고리즘은 각각의 속성 값 중 동일한 값에 대해 카운트를 하게 된다. 예를 들어, 연봉이라는 속성이 2999만원, 3000만원, 3001만원, 3002만원이라는 값을 갖는다면, 연관규칙 탐사를 위한 속성의 아이템으로 2999만원, 3000만원, 3001만원, 3002만원이라는 각각의 아이템으로 분리가 되며, 최소 지지도가 2일 때, 연봉이라는 속성으로 도출할 수 있는 연관규칙 탐사는 찾을 수가 없게 된다. 이때, 2000-2999만원은 2000만원대, 3000-3999만원은 3000만원대로 일반화 시키는 조건이 있다면, 연봉의 속성은 각각 2000만원대, 3000만원대, 3000만원대, 3000만원대로 일반화 되며 최소 지지도 2를 만족하는 3000만원대라는 빈번한 아이템을 얻을 수 있다. 따라서 본 논문에서는 사용자의 정의에 따라 비공간 속성을 일반화하는 과정을 제안한다. 사용자의 정의에 따라 일반화의 정도가 달라지기 때문에 다양한 연관규칙을 얻을 수 있다.

셋째, 공간 연관규칙 탐사 기법을 이용한 시공간 연관규칙 탐사를 위해서 시간 개념의 데이터를 비공간 속성과 동일한 방식으로 저장하고 이를 처리할 수 있는 방안을 제시한다. 공간 연관규칙 탐사 기법에 시간 개념이 추가된 시공간 연관규칙 탐사를 하기 위해서 시간 데이터는

비공간 데이터와 같이 트랜잭션 데이터베이스에 속성 값으로 저장한다. 시간 데이터의 특성상 데이터의 단위에 따라 최소 지지도와 최소 신뢰도의 값을 만족하지 못할 수 있기 때문에 시간 데이터의 단위를 일반화 시키는 과정이 필요하다. 시간 데이터의 일반화 방법은 위의 둘째 방법에서 제안한 비공간 속성 일반화 방법과 동일하다. 물론, 저장된 시간 데이터의 단위에 따라 일반화 과정이 필요하지 않을 수도 있다. 이러한 시간 데이터의 일반화 과정과 공간 데이터의 공간 술어 정의, 비공간 데이터의 일반화를 통해 얻어진 데이터를 시공간 연관규칙 탐사에 쉽게 적용할 수 있다.

제안된 방법을 통한 시공간 연관규칙 탐사의 과정은 그림 8과 같다. SD-Miner에서의 시공간 연관규칙 탐사를 위해서는 먼저 기본적으로 공간 객체의 공간 속성이 저장되어 있는 테이블, 시간 속성과 비공간 속성이 저장되어 있는 테이블, 공간 술어 개념 계층 테이블이 필요하다. 위에서 설명한 비공간 속성의 일반화를 위해 정의된 테이블이 추가로 필요하며, 연관규칙 탐사를 시도할 기준 타겟, 기준 타겟의 가장 상위 단계의 최소 지지도, 최소 신뢰도를 입력 매개변수로 받게 된다. 입력 받은 값들을 이용하여 먼저 각 공간 객체 타입의 수준에 대한 최소 지지도와 최소 신뢰도를 계산하게 된다. 그 다음, 결과가 저장될 테이블과 임시로 변환한 데이터들이 저장될 테이블이 생성되며, 이중 객체간의 공간적 술어와 비공간 속성, 시간 속성의 일반화된 값이 임시 테이블에 저장된다. 구축된 임시 테이블을 이용하여 Apriori 알고리즘을 수행하면 시공간 연관규칙 탐사 작업을 수행할 수 있다. 위의 그림 8에서 G_close_to관계 형성 단계는 공간 객체들의 공간 술어를 정의하는 단계이며, Abstract_attribute_RDB는 비공간 속성과 시간 속성을 일반화 시켜주는 단계이다. 분석된 시공간 연관규칙은 결과 테이블에 저장이 되는데, 다수의 술어부분과 다수의 결과로 규칙이 분석되어 저장된다.

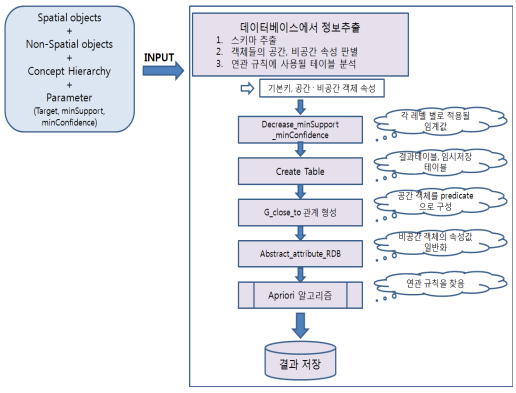

그림 8. 시공간 연관규칙 탐사의 과정

5. SD-Miner 수행 사례

본 장에서는 SD-Miner를 통해 실제 데이터를 이용한 공간 데이터 마이닝 작업 수행 사례를 제시한다. 본 사례에 쓰인 입력 데이터는 서울시 구로구의 실제 건물 데이터와 건물별 계절별 전력 사용량 데이터이다. 서울시 구로구에 존재하는 실제 건물 데이터의 좌표 값과 건물의 모양과 크기가 공간 속성으로 입력되고, 각 건물들은 그 건물의 쓰임새에 따라 타입별로 분류된다. 건물 데이터의 비공간 속성은 실제 데이터를 구하기가 어려워 건물용도 별로 생성하여 사용하였다.

SD-Mine에서는 먼저 수행하고자 하는 마이닝 함수를 선택하고(그림 9 참조) 마이닝의 대상이 되는 테이블 및 각 마이닝 함수에서 필요한 여러 가지 매개변수를 입력한다(그림 10 참조). 그림 10은 마이닝 함수 중 공간 클러스터링을 위한 매개변수 입력 예이다. 마이닝 함수에 따라 입력받아야 하는 매개변수 값은 각각 다르다.

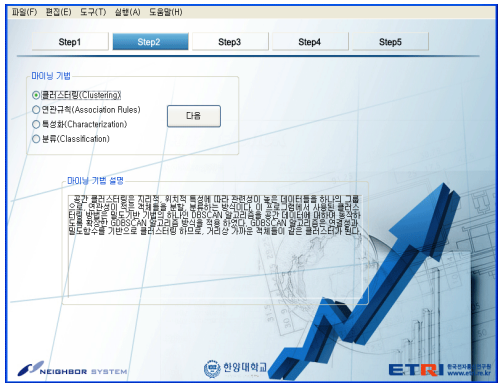

그림 9. 공간 마이닝 함수 선택 단계

그림 10-a. 테이블과 속성 선택 단계

그림 10-b. 매개변수 입력 단계

5.1 공간 클러스터링

공간 클러스터링 작업에 이용된 입력 데이터는 구로구에 위치해 있는 건물 데이터이다. 본 사례에서는 비공간 속성은 실제 값이 아니므로 공간 속성만 클러스터링 대상으로 한다.

그림 11은 사용자가 입력한 값에 대한 실제 클러스터링 결과를 보여준다. 이 결과에서는 같은 클러스터로 묶일 수 있는 건물간의 최대 거리를 0.0003(축적 0.0005), 한 클러스터로 구성될 수 있는 판단 기준을 건물 개수 30개 이상으로 설정하였다. 전체 데이터는 이 결과에서 건물들의 밀집도와 위치적 분포 특성에 따라 몇 개의 그룹으로 분할된 것을 알 수 있다(분할된 그룹은 다른 색깔로 표시). 밀집도에 따라 분할 된 건물의 그룹은 추후 다른 마이닝 알고리즘에 이용될 수 있다.

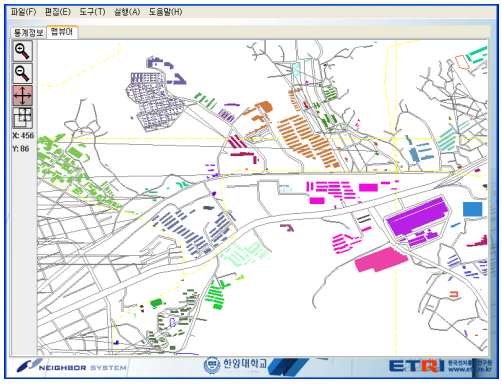

그림 11. 공간 클러스터링 결과

5.2 공간 분류

SD-Miner를 통한 공간 분류를 위한 사례로 서울시 구로구에 위치해 있는 지하철역 데이터를 모델로 다른 건물 데이터의 클래스를 예측한다. 이를 위해서 지하철역의 데이터가 트레이닝 데이터로 입력되었으며, 지하철역의 "HIGH_PROFIT"이 클래스로 사용되었다. 공간 분류를 위한 블록단위는 구로구의 구와 동 단위가 저장된 테이블이 사용되었으며, 공간 분류의 대상이 되는 데이터는 지하철역을 제외한 건물 데이터를 입력 데이터로 사용하였다.

그림 12는 공간 분류의 결과를 보여준다. 그림 12의 결과 중 "COMMUTATION PEOPLE_HIGH(N) --> HIGH_PROFIT(N)"는 "구로구의 지하철역은 통근자의 수가 많지 않으면 고수익이 아니다."라는 것을, "COMMUTATION PEOPLE_HIGH(Y) AND WORKET_ HIGH(Y) --> HIGH_PROFIT(Y)"은 "구로구의 지하철역들은 통근자의 수가 많고 일하는 사람들의 수가 많으면 고수익이다."라는 것을, "COMMUTATION PEOPLE_HIGH(Y) AND WORKER_HIGH(N) --> HIGH_PROFIT(Y)"은 "구로구의 지하철역들은 통근자의 수가 많고 일하는 사람들의 수가 많지 않아도 고수익이다."라는 것을 의미한다.

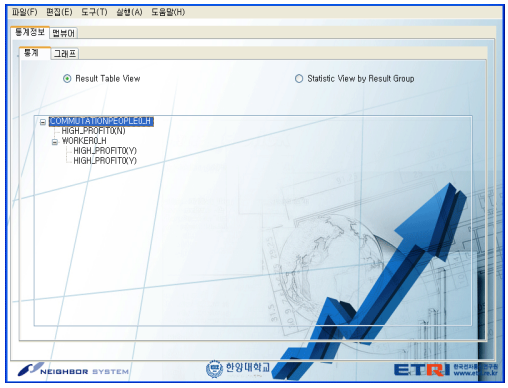

그림 12. 공간 분류 결과.

5.3 공간 특성화

SD-Miner를 통한 공간 특성화 수행 사례에서는 구로구의 건물 데이터 중 공간 클러스터링을 통해 클러스터링된 클러스터 0번 집합을 대상으로 특징을 찾아 전체 구로구 건물 데이터의 특징을 비교한다. 이를 위해서 구로구 전체의 주택 데이터를 대상으로 하며, 타겟 객체를 제 5.1절에서 수행된 클러스터링 결과 중 클러스터 0번 주택을 대상으로 공간 특성화를 수행하였다. 여기에서는 사용자가 지정한 클러스터 0번 주택들의 연간 전력 사용량을 분석하여 이웃에 위치해 있는 주택의 특성에 어떠한 영향을 미치며 전체 구로구 주택들과 어떠한 차이가 있는지 분석한다.

사용자로부터 입력된 값들을 이용하여 공간 특성화 작업을 수행한 결과 그림 13의 결과를 얻을 수 있다. 그림 13-a은 특성화 결과를 테이블로 저장하여 보여주며, 그림 13-b는 특성화된 결과를 전체 구로구 데이터를 대상으로 보여주고 그림 13-c는 특성화 된 지역만을 확대하여 보여

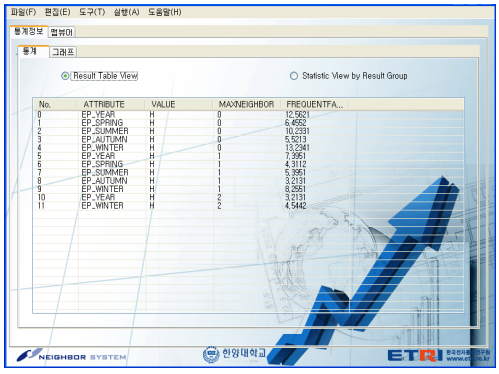

그림 13-a. 공간 특성화 결과 테이블

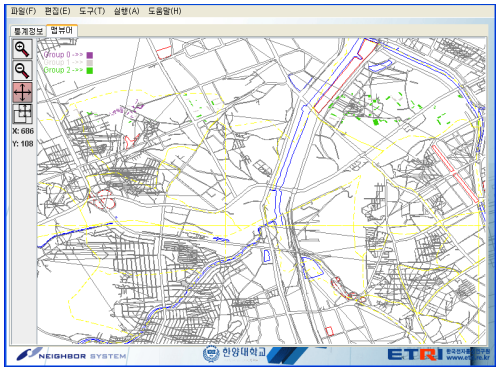

그림 13-b. 공간 특성화 결과

그림 13-c. 공간 특성화 결과 확대

준다. 그림 13-c에서 보라색으로 나타나는 그룹 0은 특성화의 대상으로 입력한 클러스터 0번에 해당되는 지역이며, 이 지역을 기준으로 한번 확장된 공간이 회색으로 표시된 그룹 1지역이다. 특성화 시 최대 이웃을 3으로 지정한 결과 초록색의 그룹 3지역까지 클러스터 0번의 특징이 확장된 것을 알 수 있다.

5.4 시공간 연관규칙 탐사

SD-Miner를 통한 시공간 연관규칙 탐사 사례를 통하여 서울시 구로구에 존재하는 건물의 종류(아파트, 전문상가, 영화관 등)와 위치에 따라 시간에 따른 전력 사용량을 탐사할 수 있다. 이를 위하여 사용된 입력 데이터는 구로구에 위치해 있는 건물 데이터이며, 각각의 건물 데이터는 건물들의 비공간 속성을 포함한다. 건물 데이터의 비공간 속성으로 각 건물들의 연간, 계절별 전력 사용량을 속성으로 포함한다.

그림 14는 도출된 연관규칙을 보여준다. 그림 14의 결과 값은 타겟이 되는 레벨1을 주택으로 놓고 최소 지지도 30, 최소 신뢰도 85%를 만족하는 연관규칙을 탐사한 것이다. 그림 14의 결과 중 “(equal,아파트) --> (EP_SUMMER00, H) (86.95%)” 의 의미는 “타겟으로 지정한 주택이 아파트이면, 여름철 전력사용량이 높다. 이것은 신뢰도 86.95%를 만족한다.” 이며, “(contains,전문상가) --> (EP_SUMMER00,H)(86.88%)”는 “타겟으로 지정한 주택에 전문상가가 포함되어있으면, 전문상가의 여름철 전력사용량이 높다. 이것은 신뢰도 86.88%를 만족한다.” 라는 것을 의미한다.

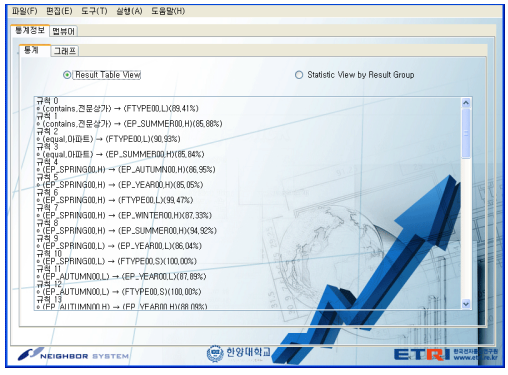

그림 14. 시공간 연관규칙 결과

6. 결 론

공간 데이터는 기존의 일반 데이터에서는 없는 거리 정보와 위상 정보를 가지고 있으며, 외형적인 구조가 데이터별로 상이하다. 또한, 공간상의 객체들은 서로 영향을 주며 존재하고 객체간의 거리나 인접성에 의해 객체들의 연관성도 객체마다 서로 다른 특징이 있다. 이러한 공간 데이터의 특징을 고려한 공간 데이터 마이닝 기법들이 연구되고 있다.

본 논문에서는 먼저, 데이터 마이닝의 주요 기법인 클러스터링, 분류, 특성화, 연관규칙 탐사에 대해 분석하고, 이를 공간적 개념으로 확장한 공간 데이터 마이닝의 개념을 분석하였다. 그리고 기존의 공간 데이터 마이닝 기법들을

분석하여 공간 데이터의 특징을 가장 잘 반영한 데이터 마이닝 기법을 채택하여 그 알고리즘의 성능을 개선하는 방안에 대해 제안하였고, 이를 이용한 공간 마이닝 툴인 SD-Miner를 설계하고 개발하였다. 또한, 개발된 SD-Miner를 이용하여 데이터를 실제 마이닝에 적용해 봄으로써, 그 결과가 유용함을 보였다.

SD-Miner에서 사용된 공간 클러스터링을 위해서는 GDBSCAN 방식을 채택하였으며 이를 더 효율적이고 정확하게 수행하기 위하여 GDBSCAN 방식을 보완하는 방법들을 제안하여 설계하고 개발하였다. 공간 분류를 위해서는 일반적인 분류 방법에서 알고리즘의 효율성을 높이기 위해 이단계 분류 방법인 RELIEF 알고리즘과 이진의 사결정트리를 사용하였다. 공간 특성화를 위해서는 기존의 특성화 방법에서 이웃관계 테이블을 추가로 이용하여 특성화의 공간적 확장을 하였다. 시공간 연관규칙 탐사를 위해서는 공간 개념 계층을 이용한 공간 연관규칙 탐사 방법에 시간 데이터를 추가로 사용하여 시공간 연관규칙 탐사를 수행하였다.

본 논문에서 제안한 SD-Miner는 일반 데이터 마이닝 함수를 확장한 개념으로, 공간 데이터 뿐 아니라, 비공간 데이터에 대해서도 마이닝이 가능하다. 사용자의 특별한 입력사항 없이 데이터의 입력 형태에 따라 시스템에서 공간 데이터인지 비공간 데이터인지를 판단하여 자동적으로 데이터 마이닝을 수행한다. 또한 각 마이닝 함수를 라이브러리 형태로 제공함으로써 각 마이닝 함수만을 따로 떼어 사용이 가능하며, SD-Miner에 다른 마이닝 함수의 추가가 용이하다. SD-Miner는 사용자의 입력을 테이블 형태로 입력받아 처리하고 공간 술어를 정의함에 있어 사용자의 의견이 반영되도록 하기 때문에 시스템의 범용성이 높다. 더불어, 알고리즘의 구현을 위해 오라클 10g에서 제공하는 함수를 이용함으로써, 알고리즘의 수행 복잡도를 줄이고, 보다 정확한 공간 데이터에 대한 여러 가지 측정값들을 얻을 수 있다. SD-Miner는 사용자로 하여금 비공간 데이터 마이닝과 더불어 공간 데이터 마이닝을 더 쉽게 다룰 수 있도록 하며, 이의 개발을 계기로 시공간 데이터 마이닝이 가능한 도구들이 더욱 발전해 나갈 것이다.

참 고 문 헌

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, *Academic Press*, 2001.

[2] W. Lu, J. Han, and B. Ooi, "Discovery of General Knowledge in Large Spatial Databases," In *Proc. Far East Workshop on Geographic Information Systems*, pp. 275-289, 1993.

[3] K. Koperski, J. Adhikary, and J. Han, "Knowledge Discovery in Spatial Databases: Progress and Challenges," In *Proc. ACM Workshop on Research Issues on Data Mining and Knowledge Discovery*, ACM SIGMOD, pp. 55-70, 1996.

[4] X. Zhou, D. Truffet, and J. Han, "Efficient Polygon Amalgamation Methods for Spatial OLAP and Spatial Data Mining," In *Proc. Int'l. Symp. on Advances in Spatial Databases*, SSD, pp. 167-187, 1999.

[5] E. Knorr and R. Ng, "Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining," *IEEE Trans. on Knowledge and Data Engineering*, IEEE TKDE, Vol. 8, pp. 884-897, 1996.

[6] J. Han, K. Koperski, and N. Stefanovic, "GeoMiner: A System Prototype for Spatial Data Mining," In *Proc. ACM Int'l. Conf. on Management of Data*, ACM SIGMOD, pp. 553-556, 1997.

[7] J. Sander et al., "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications," *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 169-194, 1998.

[8] M. Ester et al., "Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support," *Data Mining and Knowledge Discovery*, Vol. 4, pp. 193-216, 2000.

[9] M. Ester, H. Kriegel, and J. Sander, "Algorithms and Applications for Spatial Data Mining," Geographic *Data Mining and Knowledge discovery*, 2001.

[10] J. Mennis and J. Liu, "Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change," *Transactions in GIS*, Vol. 9, No. 1, pp. 5-17, 2005.

[11] 유병섭 et al., "공간 데이터 웨어하우스에서 공간 분석을 위한 공간 집계 연산," *한국공간정보시스템학회 논문지*, Vol. 9, No. 3, 2007.

[12] 엄정호 et al., "Naïve Bayesian 분류화 기법을 이용한 시간대별 평균 구간 속도 기반 주행 시간 예측 알고리즘," *한국공간정보시스템학회 논문지*, Vol 10, No. 3, 2008.

[13] SAS Co., Ltd. http://www.sas.com/technologies/analytics/datamining/miner/

[14] SPSS Co., Ltd. http://www.spss.com/clementine/

[15] IBM Co., Ltd. http://www-306.ibm.com/software/data/iminer/

[16] Rapid-i Co., Ltd. http://rapid-i.com/

[17] RuleQuest Research Co., Ltd. http://www.rule-

quest.com/
[18] NeuroDimension Co., Ltd. http://www.nd.com/neurosolutions/products/ns/whatisNN.html
[19] I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, *2nd Edition, Morgan Kaufmann*, San Francisco, 2005.
[20] I. Wang, W. Hsu, and L. M. Lee, "FlowMiner: Finding Flow Patterns in Spatio-Temporal Databases," In *Proc. Int'l Conf. on Tools with Articial Intelligence*, pp. 14-21, 2004.
[21] K. Koperski, J. Han, and N. Stefanovic, "An Efficient Two-Step Method for Classification of Spatial Data," In *Proc. Int'l. Symp. on Spatial Data Handling*, SDH, pp. 45-54, 1998.
[22] M. Ester et al., "Algorithms for Characterization and Trend Detection in Spatial Databases," In *Proc. Int'l. Conf. on Knowledge Discovery and Data Mining*, KDD, pp. 44-50, 1998.
[23] K. Koperski and J. Han, "Discovery of Spatial Association Rules in Geographic Information Databases," In *Proc. Int'l. Symp. on Advances in Spatial Databases*, SSD, pp. 47-66, 1995.
[24] M. Ester et. al., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," In *Proc. Int'l. Conf. on Knowledge Discovery and Data Mining*, KDD, pp. 226-231, 1996.
[25] Oracle Co., Ltd. http://www.oracle.com/
[26] F. Verhein and S. Chawla, "Mining Spatio-Temporal Association Rules, Sources, Sinks, Stationary Regions and Thoroughfares in Object Mobility Databases," In *Proc. Int'l. Conf. on Database Systems for Advanced Applications*, DASFAA, pp. 187-201, 2006.

배 덕 호
2006년 2월 한양대학교 정보통신대학 컴퓨터전공(학사)
2008년 2월 한양대학교 전자컴퓨터통신공학과(석사)
2008년 3월~현재 한양대학교 전자컴퓨터통신공학과(박사과정)
관심분야 : 임베디드 DBMS, 플래시 메모리 DBMS, 이동 객체 DBMS, 데이터 마이닝, 사회연결망분석

백 지 행
2005년 단국대학교 전자컴퓨터공학부(학사)
2008년 한양대학교 전자통신컴퓨터공학과(석사)
2008년~현재 (주)NHN 서비스 근무
관심분야 : 데이터 마이닝, 이동 객체 DBMS, 이동객체 관리, 텔레매틱스, 시공간 데이터베이스

오 현 교
2008년 2월 한양대학교 정보통신대학 컴퓨터전공(학사)
2008년 3월~현재 한양대학교 전자컴퓨터통신공학과(석사과정)
관심분야 : 사회 연결망 분석, 공간 데이터베이스/GIS, e-비즈니스, 데이터 마이닝

송 주 원
1981년 2월 경북대학교 공대 전자공학과 전자계산전공(학사)
1983년 2월 한국과학기술원 전산학과(석사)
1997년 8월 한국과학기술원 전산학과(박사)
1999년 5월 KAIST-KT 테크노경영MBA단기과정(연구논문보고서 최우수상)
1983년 3월~2006년 12월 KT 연구개발부문 근무(수석연구원)
2007년 9월~현재 한양대학교 BK21 사업단 정보기술분야 연구교수
관심분야 : 이동객체 데이터베이스, 다중 데이터베이스, 공간 데이터베이스/GIS, 다차원 엑세스 방법, 버전 관리 시스템, CAD 데이터베이스, 데이터마이닝

김 상 욱

1989년 2월 서울대학교 컴퓨터공학과(학사)
1991년 2월 한국과학기술원 전산학과(석사)
1994년 2월 한국과학기술원 전산학사(박사)
1991년 7월~1991년 8월 미국 Stanford University, Computer Science Department, 방문 연구원
1994년 3월~1995년 2월 KAIST 정보전자연구소 전문 연구원
1999년 8월~2000년 8월 미국 IBM T.J. Watson Research Center, Post-Doc.
1995년 3월~2003년 2월 강원대학교 정보통신공학과 부교수
2003년 3월~현재 한양대학교 정보통신대학 정보통신학부 교수
관심분야 : 데이터베이스 시스템, 저장 시스템, 트랜잭션 관리, 데이터 마이닝, 멀티미디어 정보 검색, 공간 데이터베이스/GIS, 주기억장치 데이터베이스, 이동 객체 데이터베이스/텔레매틱스, 사회 연결망 분석, 웹 데이터 분석

최 명 희

2006년 군산대학교 정보통계학과 졸업(학사)
2008년 군산대학교 대학원 정보통계학과 졸업(석사)
2007년~현재 네이버시스템㈜ 연구원
관심분야 : 시맨틱 웹, LBS, 정보보안, DB, 모바일 서비스 솔루션, 데이터마이닝

조 현 주

2000년 부산대학교 산업공학과 졸업(학사)
2000년~현재 네이버시스템㈜ 책임연구원
관심분야 : 모바일 웹2.0, 모바일 서비스 솔루션, LBS, 데이터마이닝