공간 데이터 마이닝 시스템의 설계 및 구현

백지행*, 오현교*, 배덕호*, 송주원*, 김상욱*, 최명회**, 조현주**
*한양대학교 전자컴퓨터통신공학과
**네이버 시스템(주) 모바일 사업부
e-mail: oracle@agape.hanyang.ac.kr

Design and Implementation of a Spatial Data Mining System

Ji-Haeng Baek*, Hyun-Kyo Oh*, Duck-Ho Bae*, Ju-Won Song*, Sang-Wook Kim*,
Myoung-Hoi Choi**, Hyeon-Ju Jo**
*Department of Electronics and Computer Engineering, Hanyang University
**Neighbor System Co., LTD Mobile Div.

요 약

GIS 기술의 발달로 많은 양의 공간 데이터가 축적됨에 따라 공간 데이터 마이닝의 중요성이 커지고 있다. 본 논문에서는 새로운 공간 데이터 마이닝 시스템인 SD-Miner를 제안한다. SD-Miner는 크게 GUI 모듈과 데이터 마이닝 함수 모듈, 데이터 관리 모듈의 세부분으로 구성된다. GUI 모듈은 사용자의 입력과 출력을 담당한다. SD-Miner의 핵심 부분인 데이터 마이닝 함수 모듈은 공간 데이터 마이닝의 주요 기법인 공간 클러스터링, 공간 분류, 공간 특성화, 시공간 연관규칙 탐사 기능을 제공한다. 데이터 관리 모듈은 DBMS를 이용하여 데이터를 저장하고 관리한다. 실제 공간 데이터를 이용한 마이닝을 수행함으로써 개발된 SD-Miner의 실용성을 규명하고, 의미 있는 마이닝 결과들을 도출한다.

1. 서론

최근 정보 및 컴퓨터 기술의 발달로 인해 다양한 분야에서 데이터들이 쏟아져 나오고 있다. 지리학 분야에서도 원격탐사, 지리 정보 시스템(GIS), 범지구 위치결정 시스템(GPS) 등의 발달로 인해 많은 양의 공간 데이터가 축적되고 있다. 이러한 대용량 공간 데이터 내에 잠재되어 있는 유용한 정보들과 지식들을 추출하는 마이닝 기법에 관한 연구들이 많이 시도되어 왔다[1,4,5,6,7,9,12,13,14].

그러나 공간 데이터 마이닝을 위한 기법들은 많이 연구 되었지만, 실제 공간 데이터 마이닝을 수행하기 위한 상용화된 시스템 개발은 아직 미흡한 상태이다. 현재, 공간 데이터 마이닝을 위한 시스템으로써 GeoMiner[2]가 제안되었지만, 아직 상용화되지 않았으며 입력되는 데이터의 형식에 제한이 있다. 따라서 일반 데이터 마이닝보다 더욱 전문적인 지식이 요구되는 공간 데이터 마이닝을 실제 응용에서 손쉽게 사용하기 위해서는 범용적인 공간 데이터 마이닝 시스템 개발이 필요하다.

본 논문에서는 새로운 공간 데이터 마이닝 시스템인 SD-Miner는 제안한다. SD-Miner는 공간 데이터 마이닝의 네 가지 주요 기법인 공간 클러스터링(spatial clustering)[6], 공간 분류(spatial classification)[5], 공간 특성화(spatial characterization)[1], 시공간 연관규칙 탐사(spatio-temporal association rule)[4,8,10,11] 등의 마이닝 기법을 제공한다.

SD-Miner는 공간 데이터뿐만 아니라 비 공간 데이터에 대한 마이닝도 가능하며, 각 마이닝 함수들을 라이브러리 형태로 제공하기 때문에 다른 시스템에서도 쉽게 사용 가능하다. 또한, 마이닝 매개 변수들을 테이블의 형태로 입력받기 때문에 시스템의 범용성이 높다.

마지막으로, SD-Miner의 실용성을 규명하기 위하여 실제 공간 데이터를 이용한 마이닝을 수행함으로써, 의미 있는 마이닝 결과들을 도출한다.

2. SD-Miner

SD-Miner는 크게 그래픽 사용자 인터페이스(GUI: graphic user interface) 모듈, SD-Miner 모듈, DBMS(database management system)관리 모듈의 3가지 부분으로 이루어진다.

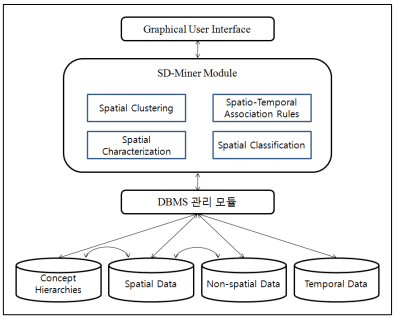

그림 1. SD-Miner 구조.

그림 1에서와 같이 GUI 모듈은 사용자로부터 마이닝에 필요한 여러 가지 매개 변수들을 입력받아 SD-Miner 모듈에 전달하고 마이닝 된 결과를 테이블이나 차트, 지도 등의 형태로 보여주는 역할을 한다. 각 마이닝 함수들의 입력 매개 변수와 결과는 조금씩 차이가 있으며, 이로 인해 각 함수별 입력 및 결과 형식이 다르게 나타난다.

SD-Miner 모듈은 GUI 모듈에서 입력받은 값들을 이용하여 마이닝을 수행하며, 마이닝된 결과를 DBMS에 전달하는 역할을 한다. 마이닝 함수로는 공간 클러스터링, 공간 분류, 공간 특성화, 시공간 연관규칙 탐사의 4가지가 제공된다.

SD-Miner에서 공간 클러스터링은 DBSCAN을 공간적으로 확장한 GDBSCAN[6]을 채택하였으며, 공간 분류는 RELIEF 알고리즘과 이진의사결정트리를 사용하여 알고리즘의 효율성을 높이는 이단계 분류 방법[5]을 채택하였다. 공간 특성화는 공간적 이웃 관계를 이용하여 특성화를 공간적으로 확장한 방법[1]을 채택하였다. 시공간 연관규칙 탐사는 Apriori 방법[3]에 공간 개념 계층과 시간 계층을 추가한 방법[4]을 채택하였다.

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

DBMS 관리 모듈은 4가지 형식의 데이터를 저장하고 사용한다. 공간 정보를 포함하는 공간 데이터, 공간 정보 외의 비공간 속성을 포함하는 비공간 데이터, 시공간 연관규칙 탐사에서 시간 개념으로 사용될 시간 데이터, 공간적 술어가 저장된 개념 계층 데이터이다.

SD-Miner의 장점은 다음과 같다. 첫째, 공간 데이터 마이닝 함수들은 공간 데이터 뿐 아니라, 비공간 데이터에 대한 마이닝도 가능하다. 이때, 사용자의 특별한 입력 사항 없이 입력된 데이터에 따라 시스템에서 공간 데이터인지 비공간 데이터인지를 자동적으로 판단하여 데이터 마이닝을 수행하게 된다.

둘째, 각 마이닝 함수를 라이브러리 형태로 제공한다. 따라서 각 마이닝 함수만을 타 시스템에서도 사용이 가능하며, 다른 마이닝 함수의 추가도 용이하다.

셋째, 사용자의 입력을 테이블 형태로 입력받아 처리하기 때문에 시스템의 범용성이 높다. SD-Miner에서는 일부 매개 변수를 제외한 모든 사용자의 입력을 테이블의 형태로 입력받는다.

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. 따라서 SD-Miner에서는 공간 술어를 정의할 때 개념 계층의 테이블을 이용하여 원하는 수준의 공간 술어를 사용자가 정의할 수 있도록 한다.

3. SD-Miner 수행사례

본 장에서는 SD-Miner를 통해 실제 공간 데이터를 이용한 공간 데이터 마이닝 작업 수행 사례를 제시한다. 본 사례에 쓰인 입력 데이터는 서울시 구로구의 건물 데이터와 건물별 계절별 전력 사용량 데이터이다.

3.1. 공간 클러스터링

본 절에서는 구로구의 건물 데이터를 이용하여 공간 클러스터링을 수행하였다. 그림 2는 클러스터링 결과를 보여준다. 클러스터링 수행 시, 클러스터로 묶일 수 있는 건물간의 최대 거리를 0.0003(축적 0.0005), 한 클러스터로 구성될 수 있는 판단 기준을 건물 개수 30개 이상으로 설정하였다. 결과에서 건물들의 밀집도와 위치적 분포 특성에 따라 몇 개의 그룹으로 분할 된 것을 알 수 있다(분할된 그룹은 다른 색깔로 표시됨).

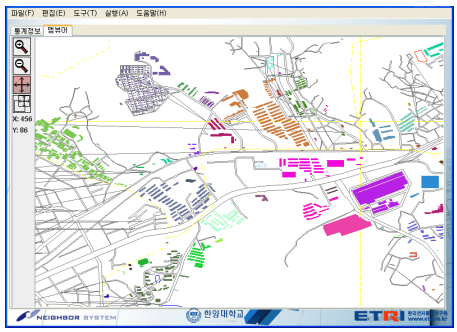

그림 2. 공간 클러스터링 결과.

3.2. 공간 분류

본 절에서는 공간 분류를 위한 사례로 서울시 구로구에 위치해 있는 지하철역 데이터를 훈련 데이터로 사용하여 지하철역을 제외한 건물 데이터의 클래스를 예측한다. 이 때, 지하철역의 비공간 속성인 "HIGH_PROFIT"이 클래스 속성으로 사용하였다.

그림 3은 공간 분류의 결과를 보여준다. 그림 3의 결과 중 "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 3. 공간 분류 결과.

그림 4(c). 공간 특성화 결과 확대.

3.3. 공간 특성화

본 절에서는 공간 특성화를 수행하였다. 이를 위해 3.1절의 클러스터들 중 클러스터 0번 집합에 해당하는 건물 데이터를 전체 구로구 건물 데이터와 비교하여 특성화한다.

그림 4는 공간 특성화 작업을 수행한 결과를 보여준다. 그림 4(a)는 특성화 결과를 테이블로 저장하여 보여주며, 그림 4(b)는 특성화 결과를 전체 구로구 데이터를 대상으로 보여주고 그림 4(c)는 특성화 된 지역만을 확대하여 보여준다. 특성화 시 최대 확장 이웃을 3으로 지정한 결과 초록색의 그룹 3지역까지 클러스터 0번의 특징이 확장된 것을 알 수 있다.

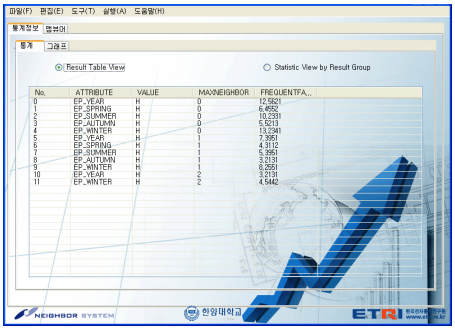

그림 4(a). 공간 특성화 결과 테이블.

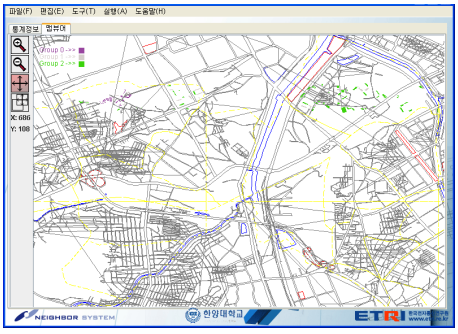

그림 4(b). 공간 특성화 결과.

3.4. 시공간 연관규칙 탐사

본 절에서는 서울시 구로구에 존재하는 건물의 종류(아파트, 전문상가, 영화관 등)와 위치에 따라 시간에 따른 전력 사용량을 탐사하였다. 이를 위하여 사용된 입력 데이터는 구로구에 위치해 있는 건물 데이터이며, 건물 데이터의 비공간 속성으로 각 건물들의 연간, 계절별 전력 사용량을 속성으로 포함한다.

그림 5는 도출된 연관규칙을 보여준다. 그림 5의 결과 중 “(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 5. 시공간 연관규칙 결과.

4. 결론

GIS 기술의 발달로 인해 많은 양의 공간 데이터가 축적됨에 따라 공간 데이터 마이닝의 중요성이 커지고 있다. 본 논문에서는 데이터 마이닝을 실제 응용에서 손쉽게 사용하기 위한 범용적인 공간 데이터 마이닝 시스템을 설계, 구현하였다.

본 논문의 공헌은 다음과 같다. 첫째, 기존의 공간 데이터 마이닝 기법들을 분석하여 공간 데이터의 특성을 가장 잘 반영한 데이터 마이닝 기법을 선택하였다. 둘째, 범용적인 공간 데이터 마이닝 시스템인 SD-Miner를 설계하고 개발하였다. 셋째, 개발된 SD-Miner를 이용하여 데

이터를 실제 마이닝에 적용함으로써, SD-Miner의 실용성을 규명하였다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었습니다. (IITA-2008-C1090-0801-0040)

참고문헌

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] J. Ale and G. Rossi, "An Approach to Discovering Temporal Association Rules," In *Proc. ACM Int'l. Symp. on Applied Computing,* ACM SAC, Vol. 1, pp. 294-300, 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," In *Proc. Int'l. Conf. on GeoComputation,* 2003.

[11] 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.

[12] 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.

[13] 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.

[14] R. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," In *Proc. ACM Int'l. Conf. on Very Large Data Bases*, pp. 144-155, 1994.