SIMS를 위한 공간 데이터 마이닝 질의 언어*

박선 박상호° 안찬민 이윤석 이주홍
인하대학교 컴퓨터정보공학과
(sunpark, parksangho°, ahnch1, aprilia)@datamining.inha.ac.kr juhong@inha.ac.kr

Spatial Data Mining Query Language for SIMS*

Sun Park Sang-ho Park° Chan-Min Ahn Youn-Seok Lee Ju-Hong Lee
School of Computer Science and Engineering, Inha University

요 약

SIMS는 공간 정보 관리 환경을 지원하기 위한 통합 관리 시스템으로서 다양한 공간 및 비공간 자료를 관리하고 여러 응용작업을 지원한다. 본 논문에서는 기존의 공간 데이터 마이닝 질의 언어가 처리하는 공간자료에 한정되지 않고, 자동 데이터 수집, 인공위성 측위 서비스, 원격탐사, GPS, 모바일 컴퓨팅 등의 다양한 자료와 시공간(Spatio-Temporal) 자료로부터 유용한 정보를 발견 할 수 있도록 SIMS를 기반으로 한 공간 데이터 마이닝 전용 시스템을 지원하는 공간 데이터 마이닝 질의 언어를 설계하였다.

1. 서 론

본 논문에서는 공간 데이터 마이닝 작업을 효율적으로 지원하기 위하여 SIMS를 기반으로 한 공간 데이터 마이닝 전용 시스템을 지원하는 공간 데이터 마이닝 질의 언어를 설계한다. 제안된 질의 언어는 Open GIS 스펙[1]에 부합하도록 확장 설계되었기 때문에 Open GIS 스펙을 지원하는 데이터베이스의 자료는 자료 변환 작업 없이 이용할 수 있고, 공간 및 비공간 자료를 처리 할 수 있는 characteristic, association, classification, cluster analysis과 SIMS의 GPS컴포넌트, RS(Remote Sensing) 컴포넌트, LDT(Location Determining Technologies)컴포넌트, ITS(Intelligent Transportation System)컴포넌트 등의 다양한 자료 및 시공간(Spatio-Temporal)자료를 처리 할 수 있도록 trend analysis를 지원한다.

2. 관련 연구

현재까지 공간 데이터 마이닝의 질의 언어에는 SQL과 같은 표준은 없으나, 공간 데이터 마이닝 시스템의 지원 요소로서 연구되고 있다. 지금까지 연구된 대표적인 공간 데이터 마이닝 질의 언어는 GeoMiner[2], GWiM[3]과 같은 공간 데이터 마이닝 전용 시스템을 지원을 위한 것과, 퍼지 공간 객체 지향 언어(Fuzzy Spatial OQL)[4], SDMOQL[5] 과 같이 공간 데이터베이스에 데이터 마이닝 작업을 지원하는 질의 언어가 있다. 위의 두가지 중에서 공간 데이터 마이닝 전용 시스템을 이용한 공간 분석이 공간 데이터베이스를 이용한 것보다 최적화된 마이닝 작업을 지원하기 때문에 효율적이다. 그러나 공간 데이터 마이닝 전용 시스템은 다른 데이터베이스의 자료를 이용하기 위해 공간 데이터 마이닝 시스템에 맞는 구조로 자료를 변환해야 하는 문제가 있다. 공간 데이터베이스는 이와는 반대로, 기존 다른 데이터베이스의 자료를 효율적으로 이용 할 수 있으나 공간 데이터 마이닝을 위한 전용 시스템에 비하여 효율이 낮다.

* 본 연구는 대학 IT연구센터 육성 지원사업의 연구결과로 수행되었음

3. SIMS의 소개

공간 정보 관리(Spatial Information Management) 환경을 지원하기 위한 SIMS는 SIMS 엔진, 외부 데이터 획득 서브시스템, 응용 서브시스템, 보안 서브시스템으로 구성되며, 응용 프로그래머는 응용 서브시스템의 인터페이스를 사용하여 공간 정보 관리 환경을 지원하는 다양한 응용 시스템을 개발 할 수 있다. SIMS 엔진은 3D 데이터 처리 컴포넌트, 공간 데이터웨어하우징 컴포넌트의 응용 서브시스템과 직접 인터페이스 하는 컴포넌트, 하이브리드 컴포넌트, 시공간 컴포넌트, P2P 컴포넌트 등의 하위 시스템 컴포넌트들로 구성된다. 또한, 응용 서브시스템의 서비스 목적에 따라 각기 다른 컴포넌트가 SIMS의 엔진 컴포넌트로 사용될 수 있다. 외부 데이터 획득 서브시스템은 외부 데이터 획득 및 추출을 위한 서브시스템이다. 외부 데이터 획득 서브시스템은 GPS 컴포넌트, RS 컴포넌트, LDT 컴포넌트, ITS 컴포넌트로 구성된다. 이들 컴포넌트에 의해 획득된 데이터는 응용 서브시스템에서 이용된다. 응용 서브시스템은 SIM 관련 응용 시스템의 구현에 사용되거나 SIMS 응용 S/W 개발 지원 서브시스템으로 공간 데이터 마이닝 컴포넌트, LBS 응용 컴포넌트로 구성된다. 보안 서브시스템은 다단계

보안 모델을 SIMS 시스템에 적용한 서브시스템으로서 컴퓨터 침입 대용 기술, 암호화 기술, 공간 데이터베이스 시스템 보안 기술 등의 정보 보호 관리 기술을 포함한다 [6].

4. 질의 언어 설계 기준

위의 관련연구와 SIMS에 처리되는 자료 분석을 통해 SIMS를 위한 공간 데이터 마이닝 질의 언어의 기능을 정의하기 위해서는 다음의 사항을 만족해야 한다.

첫째, 공간 데이터베이스의 자료를 효율적으로 처리할 수 있어야 한다.

둘째, 공간 데이터 마이닝 질의 언어를 쉽게 사용할 수 있어야 한다.

셋째, 다른 공간 데이터베이스의 자료를 이용할 수 있어야 한다.

넷째, 다양한 종류의 자료로부터 지식을 추출할 수 있어야 한다.

본 논문에서는 위의 네 가지 조건을 만족하기 위하여 다음과 같이 공간 데이터 마이닝 질의 언어를 설계하였다. 첫 번째 조건을 만족하기 위하여 SIMS를 기반으로 한 공간 데이터 마이닝 전용 시스템을 지원하는 공간 데이터 마이닝 질의 언어를 설계하였다. 두 번째와 세 번째 조건을 만족하기 위하여 Open GIS의 스펙을 지원하는 SQL언어를 공간 데이터 마이닝 질의 언어를 지원하도록 확장하였다. 이렇게 하여 SQL을 사용할 수 있는 사용자는 다른 사용법을 배우지 않고 몇 가지 옵션만으로 공간 데이터 마이닝 작업을 수행 할 수 있다. 또한 Open GIS 스펙을 지원하는 공간 데이터베이스의 자료를 자료변환 없이 이용 할 수 있는 장점과, 각각의 데이터베이스를 통합하여 관리 할 수 있도록 데이터 웨어하우스를 지원하도록 설계하였다. 마지막 조건을 만족하기 위하여 characteristic, association, classification, cluster analysis, outlier detection, trend analysis 등의 다양한 데이터 마이닝 기능을 지원하도록 설계 하였다.

5. SIMS의 공간 데이터 마이닝 질의 언어

5.1 공간 데이터 웨어하우스 조작 질의 언어

본 절에서는 데이터 웨어하우스의 스키마인 큐브를 생성하는 질의 언어에 대하여 정의하고 사용 방법에 대하여 설명한다. 공간 데이터 웨어하우스 조작 질의 언어는 DMQL[7]을 기반으로 표준 SQL을 확장하였다. 다음은 데이터베이스의 데이터 정의문을 확장하여 데이터 웨어하우스의 데이터를 정의한 것이다.

create cube <cube_name> [<dimension_list>]
<measure_list> **on** <cube_type>];
<cube_type> ::= <star> | <snowflake> | <constellation>;
create dimension <dimension_name>
[<attribute> | <subdimension_list>];

여기서 "create cube" 절은 fact 테이블을 생성하며 다루는 자료의 유형에 따라 cube 스키마를 선택한다. cube 스키마는 Stars, Snowflakes, Fcat Constellations로 구성된다. "create dimension" 절에서는 fact 테이블과 연결된 dimension 테이블을 생성한다.

다음은 데이터 웨어하우스를 제거하는 질의어이다. "drop cube"절을 이용하여 fact 테이블을 제거하며 "drop dimension"절을 이용하여 dimension 테이블을 제거한다.

drop cube <cube_name>
drop dimension (<dimension> | <subdimension>)

다음은 공간 데이터 웨어하우스를 조작하는 질의 언어이다. roll up은 dimension의 차원을 감소하거나 상위 개념 계층으로 올리는 연산을 수행하며, drill down은 roll up과는 반대의 연산을 수행한다.

roll up <attribute(s)> | <dimension>
drill down <attribute(s)> | <dimension>

5.2 공간 데이터 마이닝 조작 질의 언어

본 절에서는 SIMS의 데이터베이스 및 공간 데이터 웨어하우스로부터 사용자가 공간 데이터 마이닝과 공간 데이터 관련 의사결정을 지원할 수 있도록 공간 데이터 마이닝 질의어의 규격을 정의한다. 다음의 확장된 질의 언어에는 SQL 데이터 조작문과 유사한 공간데이터 분석을 위한 구문을 추가 하였다. 즉, "select" 절을 "mine for analyze" 절로 확장하여 공간 데이터 마이닝 작업을 지원하였으며, "from"절 이하부터는 Open GIS 의 표준 SQL 을 지원하여 사용자로 하여금 새로운 사용방법을 학습하지 않고 그대로 이용할 수 있도록 하였다.

<SMQL> ::= <SMQL_Statement>;{<SMQL_Statement>}
<SMQL_Statement> ::= <Spatial_Data_Mining_Statement>|
<Spatial_Data_Mining_Statement> ::=
mine <Kind_Of_Task> **as** <Pattern_Name>
for <Pattern_Concept>
[**analyze** <measure(s)>]
from <relation(s)> | <cube(s)>
[**where** <condition>]
[**group by** attribute, { , attribute }]
[**having** <condition(s)>]
[**set** threshold_specification **THRESHOLD** number]
<Kind_Of_Task> ::= **characteristic**|**association**|
classification|**clustering**|**trending**

"<Spatial_Data_Mining_Statement>" 에서, "mine <Kind_Of_Task> as <Pattern_Name>" 의 <Kind_Of_Task> 절은 어떤 공간 데이터 마이닝 작업을 할 것 인지 선택한다. 지원하는 데이터 마이닝 기능은 characteristic, association, classification, cluster analysis, trend analysis 이다. cluster analysis에 옵션

을 지정하면 outlier detection 작업을 수행한다. <Pattern_Name> 절은 발견되는 패턴의 이름을 지정한다. "for" 는 결과 패턴을 지정한다. <Pattern_Concept> 에는 attribute, dimension list, predicate, function 등이 지정된다. "from <relation(s)> | <cube(s)> [where <condition>]" 절은 from과 where절에 연관된 데이터베이스 테이블 이나 데이터 큐브들 그리고 데이터의 검색을 위한 조건들을 기술한다. "[group by attribute, { , attribute }]" 절은 데이터를 그룹화하기 위한 기준을 기술한다. "[having <condition(s)>]" 절은 데이터의 그룹에 관련 있는 조건을 기술한다.

5.3 공간 데이터 마이닝을 위한 Predicate

공간 데이터 마이닝은 SIMS에서 지원하는 공간 관계 연산자와 공간 연산자들을 이용하여 이루어 질 수 있다. 그러나 위에서 언급하였듯이 응용 프로그램에 따라서 특화된 공간 서술자가 필요하게 된다. 본 논문은 SIMS가 제공하는 공간 서술자 외에 거리, 방향 서술자와 시간에 관한 서술자를 추가로 정의하여 기존의 공간 마이닝의 작업 영역을 확장하고, 시계열 데이터 마이닝을 지원한다[표1].

표1) SMQL을 위한 추가 공간 Predicate

predicate	특 징	parameter	
		D	P
distanceRegion (obj1,obj2,D,P)	두 객체 사이의 "P"의 공간 관계 연산이 "D"를 만족하면 참이고 아니면 거짓이다.	mm cm km in ft yd mi nmi	equal, disjoint, intersects,touches, overlaps, crosses, within, contains, envelopedintersect, indexintersect under, over, west, east, south, north
SamePeriod(X,Y,T)	객체X,Y가 동일한 기간동안(T) 활동 여부		
Starts(X,T)	객체 X가 T시간에서 활동을 시작 여부		
Finishes(X,T)	객체 X가 T시간에서 활동을 끝냈지 여부		
Start_with(X,Y)	객체 X,Y가 동시에 활동을 시작 여부		
End_with(X,Y)	객체 X,Y가 동시에 활동을 정지 여부		
Stop(X)	객체 X가 활동기간 일시적인 정지여부		

다음은 공간 데이터 마이닝 질의 언어와 Predicate을 이용하여 데이터 마이닝 작업을 하는 예이다.

예) 경기도 과천시에서 시청을 중심으로 3km이내에 있는 주택들의 월별 가스 사용량을 기준으로 상가별로 characterize 하라.

MINE CHARACTERISTIC AS "가스사용권"
FOR 수용가요금,업종
ANALYZE 수용가.지역번호="과천시", 수용가요금.요금, 수용가요금.업종="상가",
distanceRegion(상가, 주택, **3km**, **within**)
FROM 가스시설물도

6. 결론

본 논문에서는 기존의 공간 데이터 마이닝 질의 언어가 갖고 있는 문제점을 살펴보고 이를 개선할 수 있도록 질의 언어를 설계 하였다. 본 논문에서는 SIMS 기반의 공간 데이터 마이닝 전용 시스템을 지원하는 공간 마이닝 질의 언어를 설계 하여 작업의 효율성을 유지하였으며, 제안된 질의 언어는 Open GIS 스펙에 부합하도록 확장 설계하여 Open GIS 스펙을 지원하는 데이터베이스는 자료 변환작업 없이 이용할 수 있도록 하였다. 또한, 기존의 공간 데이터 마이닝 질의 언어가 처리하는 공간 자료에 한정되지 않고, 자동 데이터 수집, 인공위성 측위 서비스, 원격탐사, GPS, 모바일 컴퓨팅 등의 다양한 자료와 시공간(Spatio-Temporal) 자료를 지원 하도록 하였으며, 이런 다양한 공간 및 비공간 자료를 처리할 수 있는 공간 데이터 마이닝 기능을 지원하도록 설계하였다.

참고문헌

[1] Open GIS Consortium, Inc. OpenGIS Simple Features Specification For SQL Revision 1.1, OpenGIS Project Document 99-049, 1999

[2] Han. J., Koperski. K., Stefanovic., N., "GeoMiner: A system prototype for spatial data mining." In Proc. ACM SIGMOD Int. Conf. on Management of Data, pages 560-563, Tucson, Arizona, 1997.

[3] Popelinsky. L., "*Knowledge Discovery in Spatial Data by Means of ILP.*" In Zytkow J.M., Quafafaou M.(eds.): Principles of Data Mining and Knowledge Discovery. Proc. of 2nd Eur. Symposium, PKDD 1998

[4] Bigolin N. M., Marsala C., "Fuzzy Spatial OQL for Fuzzy Knowledge Discovery in Database." In Z.ytkow J.M, Quafafaou M.(eds.): Principles of Data Mining and Knowledge Discovery. Proc. of 2nd Eur. Symposium, PKDD 1998

[5] Tasks Donato Malerba, Annalisa Appice, Nicola Vacca, SDMOQL : An OQL-based Data Mining Query Languagefor Map Interpretation, Proc. of the Workshop on Database Technologies for Data Mining (DTDM'02) 2002

[6] 대학IT연구센터 육성·지원사업 2003년도 수행계획서: 국제 경쟁력 강화를 위한 지능형 GIS 기술 개발 및 활용연구, 인하대학교 지능형 GIS 연구센터, 2003

[7] Han et al., "DMQL: A Data Mining Query Language for Relational Databasese." In: ACM-SIGMOD'96 Workshop on Data Mining. 1996