지역정보 관리 환경을 지원하기 위한 효율적인 질의 언어의 설계

강 성 관, 이 필 규

인하대학교 컴퓨터정보공학과
e-mail : kskk1111@empal.com , pkrhee@inha.ac.kr

Design of Efficient Query Language to support Local information administration environment

Kang Sung Kwan , Rhee Phill Kyu

Dept. of Computer Information & Engineering, Inha University

요약

SIMS는 지역 정보 관리 환경을 지원하기 위한 통합 관리 시스템으로써 다양한 지역 및 비지역 자료를 관리하고 여러 응용작업을 지원한다. 본 논문에서는 지역데이터에 대한 질의를 처리하기 위해서 기존에 개발된 지역데이터 마이닝 질의 언어에 대한 모델과 지역 서술자를 기반으로 SQL형태의 객체 질의어를 새로이 정의한다. 또한 자동 데이터 수집, 인공위성 측위 서비스, 원격탐사, GPS, 모바일 컴퓨팅 등의 다양한 자료와 시지역(Spatio-Temporal)자료로부터 유용한 정보를 발견 할 수 있도록 SIMS를 기반으로 한 지역 데이터 마이닝 질의 언어를 설계하였다.

Abstract

SIMS manages data for various spatial and non-spatial as integral management system to support space information administration environment and support several application works. Without being limited to spatial data that existent spatial Data Mining question language advances handling in this paper, did so that can find useful information from various data connected with automatically data collection, artificial satellite side upside service, remote sensing, GPS. Mobile Computing and data about Spatio-Temporal. Also, we designed spatial Data Mining query language that support a spatial Data Mining exclusive use system based on SIMS.

키워드 : SIMS, 지역정보 관리,지역 데이터,Open GIS

1. 서 론

지역 데이터 마이닝은 대량의 지역 데이로부터 관심있는 부분, 특성, 관계 등을 찾아내는 기법으로 데이터 마이닝이 지역 데이터를 분석하는 분야로 확대된 것이다. 일반적인 데이터 마이닝과의 차이는 지역 데이터의 특성을 고려한 분석을 수행한다. 개념적으로 지역 데이터 마이닝을 이용할 경우, 지역 데이터 내에 포함된 데이터 간의 위상, 거리, 방향 관계를 분석의 대상으로 할 수 있으며, 데이터의 지역적 의존성(Spatial Dependency)과 이질성(Spatial Heterogeneity)을 고려하여 대용량 지역 데이터로부터 일정한 패턴과 규칙을 발견할 수 있다[1].

본 논문에서는 지역 데이터 마이닝 작업을 효율적으로 지원하기 위하여 SIMS를 기반으로 한 지역 데이터 마이닝 전용 시스템을 지원하는 지역 데이터 마이닝 질의 언어를 설계한다. 제안된 질의 언어는 Open GIS[2] 스펙에 부합하도록 확장 설계되었기 때문에 Open GIS 스펙[3]을 지원하는 데이터베이스의 자료는 자료 변환 작업없이 이용할 수 있고, 지역 및 비지역 자료를 처리할 수 있는 특징, 관계, 분류, 클러스터 분석과 SIMS의 GPS 컴포넌트, RS(Remote Sensing)컴포넌트, LDT(Location Detemining Technologies)컴포넌트, ITS(Intelligent

Transportation System)컴포넌트 등의 다양한 자료 및 시지역(Spatio-Temporal)자료를 처리할 수 있도록 경향 분석을 지원한다[2][3].

본 논문의 내용은 다음과 같이 구성되어 있다. 제 2장에서는 기존 지역 데이터 마이닝 질의 언어에 대하여 기술하고 제 3장에서는 지역 정보 관리 환경을 지원하는 SIMS의 구조에 대하여 기술한다. 제 4장에서는 SIMS를 위한 지역 데이터 마이닝 질의 언어의 설계기준에 대하여 기술하며, 제 5장에서는 이러한 질의 언어의 구체적인 종류 및 형태에 대하여 기술하고 제 6장에서 결론을 맺는다.

2. 관련 연구

지역 데이터베이스 시스템에서는 각 지역 객체의 식별자와 정확히 정의된 범위 그리고 각 지역 객체간의 지역적인 관계를 표현할 수 있어야 한다. 따라서, 효과적인 질의를 위해서는 지역 데이터베이스 안에 최적화 된 지역 데이터 모델을 정의하고 지역데이터 타입을 반드시 제공하여야 한다. 뿐만 아니라, 이러한 지역 데이터는 실질적인 데이터인 비지역 데이터와 연결 처리되어야 한다[4].

현재까지 지역 데이터 마이닝 질의 언어에는 SQL과 같은 표준은 없으나, 지역 데이터 마이닝 시스템의 지원요소로서 연구되고 있다. 지금까지 연구된 대표적인 지역 데이터 마이닝 질의 언어에는 GeoMiner, GWiM와 같이 지역 데이터 마이닝 전용 시스템을 지원하기 위한 것과, 퍼지 지역 객체 지향 언어(Fuzzy Spatial OQL)[5], SDMOQL[6]과 같이 지역 데이터베이스에 데이터 마이닝 작업을 지원하기 위한 질의 언어가 있다. 이중에서 지역 데이터 마이닝 전용 시스템을 사용하는 것이 지역 데이터베이스를 이용하여 지역 분석하는 것보다 최적화된 마이닝 방법을 지원하기 때문에 효율적이다[7].

그러나 지역 데이터 마이닝 전용 시스템은 다른 데이터베이스의 자료를 이용하기 위해 지역 데이터 마이닝 시스템에 맞는 구조로 자료를 변환해야 하는 문제가 있다. 지역 데이터베이스는 이와는 반대로, 기존 다른 데이터베이스의 자료를 효율적으로 재이용 할 수 있으나 지역 데이터 마이닝을 위한 전용 시스템에 비하여 효율이 낮다[8].

3. SIMS의 소개

지역 정보 관리(Spatial Information Management)환경을 지원하기 위한 SIMS는 SIMS 엔진, 외부 데이터 획득 서브 시스템, 응용 서브 시스템, 보안 서브 시스템으로 구성되며, 응용 프로그래머는 응용 서브시스템의 인터페이스를 사용하여 지역 정보 관리 환경을 지원하는 다양한 응용 시스템을 개발 할 수 있다.

SIMS 엔진은 3D 데이터 처리 컴포넌트, 지역 데이터웨어 하우징 컴포넌트의 응용 서브 시스템과 직접 인터페이스 하는 컴포넌트, 하이브리드 컴포넌트, 시지역 컴포넌트, P2P컴포넌트 등의 하위 시스템 컴포넌트들로 구성된다. 클라이언트가 응용 서브 시스템에 질의를 요청하고 SIMS Engine에서는 결과를 참조할 수 있는 인터페이스를 제공한다. 클라이언트는 이 인터페이스를 통하여 지역데이터 마이닝을 지원하는 SQL 문장을 실행할 수 있다.

응용 서브 시스템에서는 외부데이터 획득 서브 시스템과 연계되어 실제 데이터를 접근하는 컴포넌트를 제공한다. 이 시스템에서는 질의 처리시 필요한 지역 객체에 대한 위치정보, 속성정보의 메타정보를 관리하고 속성 조건, 시간 조건, 지역객체 ID정보 등을 이용하여 외부 데이터 획득 서브 시스템에서 지역객체 위치정보를 읽어와서 질의를 처리하기 위한 자료구조로 변환한다. 그리고 지역객체의 위치정보와 속성정보를 연결하여 하나의 객체로 만들어주는 역할을 담당하게 된다.

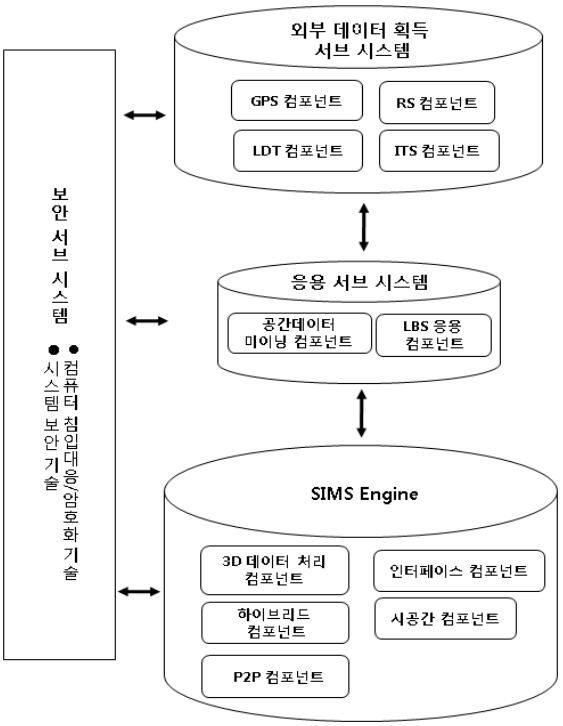

그림 1. SIMS 시스템 구조도

또한, 응용 서브 시스템의 서비스 목적에 따라 각기 다른 컴포넌트가 SIMS의 엔진 컴포넌트로 사용될 수 있다. 외부 데이터 획득 서브 시스템은 GPS 컴포넌트, RS 컴포넌트, LDT컴포넌트, ITS컴포넌트 구성된다. 이들 컴포넌트에 의해 획득된 데이터는 응용 서브시스템에서 이용된다. 응용 서브시스템은 SIM관련 응용 시스템의 구현에 사용되거나 SIMS 응용 S/W 개발 지원 서브 시스템으로 지역 데이터 마이닝 컴포넌트, LBS 응용 컴포넌트로 구성된다. 보안 서브시스템은 다단계 보안 모델을 SIMS 시스템에 적용한 서브시스템으로서 컴퓨터 침입 대응 기술, 암호화 기술, 지역 데이터 베이스 시스템 보안 기술 등의 정보 보호 관리 기술을 포함한다[9].

4. 질의 언어 설계 기준

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

첫째, 지역 데이터베이스의 자료를 효율적으로 처리할 수 있어야 한다.
둘째, 지역 데이터 마이닝 질의 언어를 쉽게 사용할 수 있어야 한다.
셋째, 다른 지역 데이터베이스의 자료를 이용할 수 있어야 한다.
넷째, 다양한 종류의 자료로부터 지식을 추출할 수 있어야 한다.

본 논문에서는 위의 네가지 조건을 만족하기 위하여 다음과 같이 지역 데이터 마이닝 질의 언어를 설계하였다. 첫 번째 조건을 만족하기 위하여 SIMS를 기반으로 한 지역 데이터 마이닝 전용 시스템을 지원하는 지역 데이터 마이닝 질의 언어를 설계하였다. 두 번째와 세 번째 조건을 만족하기 위하여 Open GIS의 스펙을 지원하는 SQL 언어를 지역 데이터 마이닝 질의 언어를 지원하도록 확장하였다. 이렇게 하여 SQL을 사용할 수 있는 사용자는 다른 사용법을 배우지 않고 몇가지 옵션만으로 지역 데이터 마이닝 작업을 수행할 수 있다. 또한 Open GIS 스펙을 지원하는 지역 데이터 베이스의 자료를 자료변환 없이 이용할 수 있는 장점과, 각각의 데이터베이스를 통합하여 관리할 수 있도록 데이터 웨어 하우스도 조작할 수 있는 질의 언어도 지원하도록 설계하였다. 마지막 조건을 만족하기 위하여 특성, 관계, 분류, 클러스터 분석, 특이한 값 검출, 경향 분석 등의 다양한 데이터 마이닝 기능을 지원하도록 설계하였다.

5. SIMS의 지역 데이터 마이닝 질의 언어

5.1 지역 데이터 웨어하우스 조작 질의 언어

본 절에서는 데이터웨어 하우스의 스키마인 큐브를 생성하는 질의 언어에 대하여 정의하고 사용 방법에 대하여 설명한다. 지역 데이터 웨어하우스 조작 질의 언어는 DMQL[10]을 기반으로 표준 SQL을 확장하였다. [표1]은 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 cube스키마를 생성한다. 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 fact테이블을 제거하며 ”drop dimension”절을 이용하여 dimension 테이블을 제거한다.

```
drop cube<cube_name>
drop dimension(<dimension> | <subdimensiion> )
```

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

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

5.2 지역 데이터 마이닝 조작 질의 언어

본 절에서는 SIMS의 데이터베이스 및 지역 데이터 웨어하우스로부터 사용자가 지역 데이터 마이닝과 지역 데이터 관련 의사 결정을 지원할 수 있도록 지역 데이터 마이닝 질의 언어의 규격을 정의한다. 다음은 질의 언어를 확장 BNF로 정의하였다. 확장된 질의 언어에는 SQL의 데이터 조작문과 유사한 지역데이터 분석을 위한 구문을 추가하였다. 즉, “select”절을 ”mine for analyze”절로 확장하여 지역 데이터 마이닝 작업을 지원하였으며, “from”절 이하부터는 Open GIS의 표준 SQL을 지원하여 사용자로 하여금 새로운 사용방법을 학습하지 않고 그대로 이용할 수 있도록 하였다. 또한 Open GIS명세서를 지원하는 데이터베이스의 자료는 자료변환 없이 곧바로 사용할 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>절은 어떤 지역 데이터 마이닝 작업을 할 것인지 선택한다. 지원하는 데이터 마이닝 기능은 characteristic, association, classification,

cluster analysis, trend analysis이다. cluster analysis에 옵션을 지정하면 outlier detection 작업을 수행한다.

<Pattern_Name>절은 발견되는 패턴의 이름을 지정하고 “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 테이블이나 데이터 큐브들 그리고 데이터의 검색을 위한 조건들을 기술한다. “[group by attribute, { , attribute}]″ 절은 데이터를 그룹화하기 위한 기준을 기술한다. ”[having <condition(s)>]″ 절은 데이터의 그룹에 관련있는 조건을 기술한다.

5.3 지역 데이터 마이닝을 위한 Predicate

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

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” AND
수용가요금.요금 AND 수용가요금.업종=“상가” AND distanceRegion (상가, 주택, 3, KM)
FROM 가스시설물도

[표 1] SMQL을 위한 추가 지역 서술자

predicate	특징	parameter	
		D	P
distanceRegion(obj1,obj2,D,P)	두 객체 사이의 “P″의 지역 관계 연산이 ”D″를 만족하면 참이고 아니면 거짓이다.	mm cm km in ft yd	equal,disjoint,intersects,touches, overlaps,crosses, within,contains, envelopedintersect, indexintersect
		mi nmi	under, over, west, east, south, north
SamePeriod (X,Y,T)	객체 X,Y가 동일한 기간동안(T) 활동하였는가?		
Starts(X,T)	객체 X가 T시간에서 활동을 시작하였는가?		
Finishes (X,T)	객체 X가 T시간에서 활동을 끝냈는가?		
Start_with (X,Y)	객체 X,Y가 동시에 활동을 시작하였는가?		
End_with (X,Y)	객체 X,Y가 동시에 활동을 정지하였는가?		
Stop(X)	객체 X가 활동기간동안 일시적인 정지를 하였는가?		

6. 결 론

본 논문에서는 기존의 지역 데이터 마이닝 질의 언어가 갖고 있는 문제점을 살펴보고 이를 개선할 수 있도록 질의 언어를 설계하였다. 본 논문에서는 SIMS 기반의 지역 데이터 마이닝 전용 시스템을 지원하는 지역 마이닝 질의 언어를 설계하여 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 지역 데이터 마이닝 질의 언어가 처리하는 지역 자료에 한정되지 않고, 자동데이터 수집, 인공위성 측위 서비스, 원격탐사, GPS, 모바일 컴퓨팅 등의 다양한 자료와 시지역(Spatio-Temporal) 자료를 지원하도록 하였으며, 이런 다양한 지역 및 비지역 자료를 처리할 수 있는 다양한 데이터 마이닝 기능을 지원하도록 설계하였다.

참고문헌

[1] Dunhan M.H. , Data Mining:Introductory and Advanced Topics. Prentice Hall, 2003

[2] J. Han, K. Koperski, and N. Stefanovic. 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] Thomas G. Dietterich, ″Machine Learning Research: Four Current Directions″, The Ai Magazine, vol 18, n0.4, 97-136, 1998

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

[5] Popelinsky L.: Knowledge Discovery in Spatial Data by Means of ILP. In Zytkow J.M., Quafafaou M.(eds.): Principles of Data Mining and Knoeledge Discovery.Proc. of 2nd Eur. Symposium, PKDD′98, Nantes, France, LNCS 1510, Springer Verlag 1998.

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

[7] Bigolin N.M., Marsala C.Fuzzzy 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 ′98, Nantes, France. LNCS 1510, Springer Verlag 1998

[8] Xiaohui Yuan, Mining Negative Association Rules, International Symposium on Computer and Communications, pp.623-628, 2005

[9] Han et al. DMQL : A Data Mining Query Language for Relational Database. In:ACM-SIGMOD'96 Workshop on Data Mining.

[10]M.Egenhofer, Spatial SQL: A query and presentation language. IEEE Transactions on Knowledge and data engineering. 6:86-95, 1994