공간 데이터웨어하우스에서 효율적인 공간 데이터 적재를 위한 이기종 데이터 소스의 비중복 추출기법

(Non Duplicated Extract Method of Heterogeneous Data Sources for Efficient Spatial Data Load in Spatial Data Warehouse)

이 동 욱[*] 백 성 하[*] 김 경 배[**] 배 해 영[***]
(Dong Wook Lee) (Sung Ha Baek) (Gyoung Bae Kim) (Hae Young Bae)

요 약 공간 데이터웨어하우스는 공간 DBMS 또는 다양한 소스데이터로부터 시간에 따라 추출된 공간데이터를 ETL 과정을 통해 가공된 데이터를 관리하는 시스템이다. 적재 주기 마다 추출된 공간데이터는 비공간 데이터와 달리 같은 주제의 중복된 공간 정보가 유용하지 않으며, 공간 데이터의 특징으로 저장 공간의 낭비가 크다. 또한 이기종간의 시스템에서 소스 데이터를 추출할 경우 서로 다른 공간데이터 타입 및 스키마를 가지고 있어 이를 위한 공간데이터 추출 기법이 요구된다. 기존 기법에서는 기준이 되는 Geocoding DB를 이용하여 추출된 공간데이터에 대한 주소 매칭과정을 수행함으로써, 정형화된 데이터 셋을 적재한다. 하지만 이 기법은 추출 데이터를 매번 Geocoding DB와 비교 연산이 발생하며, 주제별로 공간데이터를 통합 관리함에 따라 이 기종 공간 DBMS 사이에 중복된 데이터를 고려하지 않는 문제점이 있다. 본 논문에서는 공간 데이터웨어하우스 구축기 내에서 이 기종의 소스 시스템으로부터 추출된 갱신질의 통합을 이용한 효율적 추출 기법을 제안한다. 이는 이 기종의 공간 DBMS로부터 발생한 과거 적재 시점부터 현재까지 발생한 질의 중 삽입이나 삭제 등의 업데이트 관련 질의만을 추출하여 공간데이터의 불필요한 추출 연산 비용을 제거한다. 또한 소스 공간 데이터베이스 관리시스템의 업데이트 질의를 이용하여 추출된 공간데이터를 주제별로 중복 제거 및 통합 한다. 제안 기법은 데이터 중복 저장에 의한 저장 공간의 낭비를 줄이고, 적재시점 별 통합된 데이터를 적재함으로써 빠른 공간데이터 분석을 지원할 수 있다.

키워드 : 이기종 공간 데이터 추출, 비중복 추출, 공간 데이터 웨어하우스

Abstract Spatial data warehouses are a system managing manufactured data through ETL step with extracted spatial data from spatial DBMS or various data sources. In load period, duplicated spatial data in the same subject are not useful in extracted spatial data dislike aspatial data and waste the storage space by the feature of spatial data. Also, in case of extracting source data on heterogeneous system, as those have different spatial type and schema, the spatial extract method is required for them. Processing a step matching address about extracted spatial data using a standard Geocoding DB, the exiting methods load formal data set. However, the methods cause the comparison operation of extracted data with Geocoding DB, and according to integrate spatial data by subject it has problems which do not consider duplicated data among heterogeneous spatial DBMS. This paper proposes efficient extracting method to integrate update query extracted from heterogeneous source systems in data warehouse constructer. The method eliminates unnecessary extracting operation cost to choose related update queries like insertion or deletion on queries generated from loading to current point. Also, we eliminate and integrate extracted spatial data using update query in source spatial DBMS. The proposed method can reduce wasting storage space caused by duplicate storage and support rapidly analyzing spatial data by loading integrated data per loading point.

Keywords : Heterogeneous Spatial Data Extract, Non Duplicated Extract, Spatial Data Wa

† 본 연구는 건설교통부 첨단도시기술개발사업- 지능형국토정보기술혁신 사업과제의 연구비지원(07국토정보C05)에 의해 수행되었음.

[*] 인하대학교 정보공학과 박사과정, dwlee@dblab.inha.ac.kr, shbaek@dblab.inha.ac.kr
[**] 서원대학교 컴퓨터교육과 조교수, gbkim@seowon.ac.kr(교신저자)
[***] 인하대학교 정보공학부 교수, hybae@inha.ac.kr

논문접수 : 2009.03.30 수정일 : 1차 2009.05.20 / 2차 2009.05.28 심사완료 : 2009.05.30

1. 서 론

기업들의 의사결정 지원을 위해 방대한 양의 축적된 데이터를 이용하여 이를 분석할 수 있는 데이터웨어하우스에 대한 수요가 증가하고 있다. 이는 기존의 비공간데이터에 대한 기업의 의사결정을 지원하는 데이터웨어하우스 시스템 뿐만 아니라, 물류 관리, 입지 선정 등과 같이 공간정보에 대한 의사결정을 지원하기 위한 공간 데이터웨어하우스 시스템이 요구되고 있다[1, 2, 3].

공간 데이터웨어하우스 시스템은 지리정보 시스템, 교통관리 시스템 그리고 물류관리 시스템과 같이 다양한 서비스를 지원하는 이 기종의 공간 DBMS로부터 공간 및 비공간 데이터를 추출하고, 가공 및 정제하여 ODS (Operational Data Store)에 적재한다. ODS로 적재된 공간데이터는 적재 주기에 따라 공간 데이터웨어하우스로 저장되고, 공간 데이터웨어하우스는 이러한 과정을 통해 저장된 대용량의 데이터와 OLAP(On-Line Analytical Processing) 연산을 제공하여 의사결정을 지원하는 시스템이다[4, 5, 6, 7, 8, 9]. 특히 공간 데이터웨어하우스 시스템에서는 이 기종 시스템으로부터 추출되는 공간 데이터가 비공간 데이터와는 달리 시간에 따른 중복 데이터가 유용하지 않은 경우가 대부분이고, 그 크기가 대용량인 경우가 많다. 예를 들어 도로 정보를 추출한다면 새로 건설되거나 변경 된 도로의 정보는 유용하나, 변화가 없는 도로 데이터를 중복적으로 공간 데이터웨어하우스에 저장하는 것은 의미를 갖지 못하고 저장 비용만 낭비된다[10, 11]. 따라서 공간 데이터웨어하우스 시스템에서는 데이터 적재를 위한 ETL(Extract, Transform, Loading) 과정에서 공간 데이터의 특성을 고려하여 공간 계층 또는 주제별 통합 및 중복제거 기법들이 필요하다[12, 13, 14, 15, 16].

본 논문에서는 공간 데이터웨어하우스에서 효율적인 공간 데이터 적재를 위한 이기종 간의 비중복 추출 기법을 제안한다. 제안기법은 소스 시스템으로부터 과거 적재 시점부터 현재까지 발생한 질의 중 삽입이나 삭제 등 데이터의 갱신이 발생되는 질의만을 추출하여 불필요한 데이터의 추출 연산 비용을 제거한다. 또한 추출되어 임시 저장된 갱신질의에 대해서도 중복된 질의에 대한 단일질의로의 통합연산을 수행함으로써 빈번하게 갱신이 일어나는 데이터의 추출비용을 줄인다. 이러한 갱신질의의 전처리 단계를 통해 추출된 공간데이터는 주제별 중복 제거 과정 및 통합 과정을 수행함으로써 주제별 통합된 하나의 데이터만을 적재하는 과정을 수행하며, 이때 과거 적재시점에 시스템 서버에 적재된 이력 데이터와의 중복 제거연산을 통해 현시점에 발생된 갱신된 데이터만을 ODS에 적재하게 된다[17].

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로 기존의 데이터 추출 기법들을 설명을 하고, 3장에서는 제안기법에 대해 설명한다. 4장에서는 과거 적재기법과 제안기법과의 성능평가 결과를 제시한 후, 5장에서는 결론 및 향후 연구에 대해 기술한다.

2. 관련 연구

본 장에서는 기존의 데이터웨어하우스 시스템에서 제시된 이기종 소스의 추출 기법 중 시맨틱 기반의 데이터 추출 기법 및 Geocoding DB를 이용한 추출 기법, ETL 최적화를 통한 추출 기법과 비교한다.

2.1 시맨틱 기반 소스 데이터 추출 기법

데이터웨어하우스 구축을 위해 수행하는 연산 중 가장 중요한 것 중 하나는 현존하는 데이터 소스의 구조나 내용들을 데이터 손실 없이 일반적인 데이터 모델에 매핑시키는 것이다. 즉 소스 시스템에서 추출된 데이터를 정형화된 형태로 변환하여 데이터웨어하우스로 적재한다[13, 14]

이러한 적재 기법 중 유사 시맨틱 기반의 데이터 변환을 통한 적재 기법이 제안되었다[18]. 시맨틱 기반의 소스 데이터 추출 기법은 이기종 소스 시스템으로 추출된 다른 이름을 가진 데이터의 스키마를 다루기 위해 유사 시맨틱을 기반으로 공통의 사전(Dictionary)을 생성하는 어플리케이션 사전의 생성 단계 및 데이터 소스들에 주석을 달기 위한 데이터 저장소의 주석 생성 단계를 지닌다. 또한, W3C에서 지정한 OWL(Web Ontology Language)를 이용하여 이 기종이기 때문에 가지는 소스 시스템 사이의 스키마의 불일치성 즉, 추출된 소스 데이터들의 의미론적 충돌을 해결하며 이러한 의미론적 충돌 해결을 통해 추출된 데이터가 가지고 있는 이 기종간 정보들의 통합을 수행하게 된다. 추출된 데이터는 데이터웨어하우스에 적재되기 전에 스키마에 대한 매핑 과정이 필요하다. 즉 데이터웨어하우스 내부에 정의된 스키마와 일치시키는 정형화 단계가 필요하다.

하지만 공간 데이터에서 지형 또는 건축물 등의 변화가 없는 경우 중복된 데이터가 반복적으로 추출되는데, 이를 중복하여 추출 및 적재하는 것은 의미가 없으며, 시스템의 추출 및 관리 비용을 증가 시킨다.

2.2 Geocoding DB를 통한 공간데이터 적재기법

공간 데이터웨어하우스 구축을 위한 공간데이터 적재 기법으로 Geocoding DB를 이용하는 Addressing Matching 단계를 가진 이 기종의 소스 데이터 적재 기법이 제안되었다[19].

Geocoding DB를 통한 이러한 기법은 먼저 정제 단계에서 이 기종의 다른 시스템이 가지고 있는 다양한 포맷에 대한 정규화 작업을 하고, 결과로 나온 데이터에 대하여 모든 공간데이터의 좌표 값의 기준이 되는 데이터를 가지고 있는 Geocoding DB를 참조하는 Address Matching을 수행함으로써 추출된 데이터를 정규화 된 좌표 값을 가진 데이터로 변환시킨다. 그리고 이렇게 생성된 이 기종 소스 시스템의 주제별 데이터는 공간 데이터웨어하우스 시스템에서 사용되는 정규화 된 형태로 적재되게 된다. 그러나 이러한 공간데이터의 적재방법은 몇 가지 문제점을 지니고 있다.

첫째, 기준이 되는 Geocoding DB를 이용하는데 있어

서 Geocoding DB는 언제나 최신의 데이터를 가지고 있을 것이 요구된다. 즉, 공간데이터웨어하우스의 소스로 사용되는 각 시스템들로부터 얻어온 가장 최근에 업데이트가 발생된 공간데이터를 적재하는 과정에서 만약, 기준이 되는 Geocoding DB가 최신의 공간데이터에 대한 정보가 갱신 전이라면 소스 시스템으로부터 추출된 최신의 업데이트 데이터는 잘못된 데이터로 인식하여 SDW(Spatial Data Warehouse) 시스템 내에 적재되지 않는다.

둘째, Address Matching으로 정형화된 Geometric 좌표 값을 얻는 것은 이 기종 시스템으로부터 추출된 주제별 공간데이터에 대하여 하나씩 Geocoding DB와 비교연산을 수행해야 하므로 자연히 데이터 비교연산 비용증가에 따른 전체 시스템에서의 공간데이터 적재비용이 증가하는 단점이 존재한다. 또한, 같은 주제의 공간 데이터에 대해 과거 적재시점의 공간데이터와 중복 제거를 고려하지 않기 때문에 저장 공간을 낭비한다.

2.3 ETL 최적화를 통한 데이터 추출기법

기존의 데이터웨어하우스의 연구들은 ETL 과정 중 변환연산에 관련된 부분만을 주로 다루고 있었으나, 최근에는 데이터 적재과정에서의 논리적 최적화에 따른 데이터 적재방법에 대한 연구가 진행되었다[13]. 이 기법은 소스 시스템으로부터 추출한 데이터에 대하여 데이터웨어하우스로 적재 시 데이터의 형 변환과 적재과정에서의 최적화 과정을 지닌다. 이 과정에서의 적재과정의 최적화는 기존의 질의 최적화를 확장하여 전체 데이터웨어하우스 시스템에서의 데이터 적재비용을 최소화한다.

이 기법은 기존의 데이터 추출 및 적재 과정을 세분화하여 혹시나 발생할지 모르는 데이터 적재과정에서의 불필요한 데이터의 연산을 제거한다는 장점이 있다. 하지만 이러한 데이터의 적재과정 역시 앞에서 제시한 같은 주제의 데이터에 대한 중복연산의 미 지원 및 참여 소스 시스템의 개수에 따른 조인연산 비용이 증가한다는 단점이 존재한다.

3. 효율적인 공간 데이터 적재를 위한 이 기종간의 비중복 추출 기법

본 장에서는 효율적인 공간데이터 적재를 위한 이기종 시스템간의 추출기법을 제안한다. 이는 소스 공간 DBMS로부터 업데이트가 발생된 공간데이터만을 추출하고 적재시점별, 주제별로 통합하는 연산과정을 포함한다. 이러한 연산결과로 생성되어 주제별로 통합된 공간데이터는 SDW의 저장 공간의 낭비를 줄일 수 있으며, 적재 주기에 맞춰 SDW에 적재되기 때문에 시간의 흐름에 따른 공간데이터 분석을 지원할 수 있는 구조를 가진다.

이를 위해 먼저 소스로부터 공간데이터 추출을 위한 갱신질의 통합 과정에 대해 설명하고 이 과정에 의해 추출된 공간데이터에 대한 주제별 중복 제거 기법에 대해 설명한다.

3.1 이 기종 시스템에서의 갱신질의 통합

제안기법은 이 기종 소스 시스템으로부터 추출된 공간데이터와 공간 데이터웨어하우스에 기 적재된 데이터와의 비교연산을 수행함으로써 데이터의 비 중복성을 유지시킨다. 이 기종 시스템에서의 공간데이터 추출은 주제별 공간데이터의 속성에 따라 추출 주기가 다르다. 즉, 도로 또는 강과 같은 지리 정보에 관한 공간데이터는 빈번하게 업데이트가 발생하지 않아 업데이트에 따른 데이터 추출 주기가 길다. 그러나 태풍의 영향권, 여름철 장마기간 강 유역의 변화 그리고 이동체 경로를 저장하기 위한 공간데이터는 빈번하게 갱신이 발생하여 이를 추출하기 위한 비용이 크다. 또한, 추출된 공간 데이터에 대해 기 적재 데이터와 비교 연산이 수행된다.

제안기법은 공간데이터 추출과정에서 각 소스 시스템의 Insert, Delete, Update 질의와 같이 공간데이터에 대한 갱신을 발생하는 질의들에 대해 트리거를 이용하여 독립적으로 관리한다. 즉, 소스 시스템으로부터 공간데이터 추출 시 각 소스 시스템 별로 수행된 질의 중 갱신 질의들을 수집하고, 수집된 질의들을 분석하여 변동된 공간데이터를 추출한다. 이것은 추출 주기마다 모든 소스 시스템의 데이터를 추출하지 않아 시스템 비용을 줄일 수 있는 장점이 있다. [그림 1]은 갱신질의 추출을 위한 트리거 예이다.

```
CREATE TRIGGER UpdateTrigger
AFTER UPDATE OF HanRiver
REFERENCING OLD AS OldTuple, NEW AS NewTuple
WHEN ( (OldTuple .HanRiver != NewTuple .HanRiver)  AND
(OldTypleHanRiver .Time < NewTuple .HanRiver) )
UPDATE HanRiver
SET HanRiver = NewTuple .HanRiver
FOR EACH ROW;
```

그림 1. 갱신질의 추출을 위한 트리거의 예

각 소스 시스템에서 발생한 사용자 질의 중 검색질의를 제외한 업데이트 관련 질의에 대해서만 트리거를 통하여 ODS내의 임시 버퍼 공간에 저장한다. [그림 2]에서 'System 1'에서 발생한 질의는 Select 질의와 Insert 질의 중 '한강'이라는 공간데이터의 Insert 질의만을 추출하여 ODS내의 임시 저장 공간에 해당 질의를 저장한다.

[그림 2]와 같이 ODS에 수집된 일련의 갱신 질의들은 Select 질의들로 변환되어 소스 시스템으로부터의 공간데이터 추출에 사용된다.

소스 시스템으로부터 발생된 갱신 질의에 대해서만 데이터 추출을 하는 것은 모든 질의에 대해 갱신이 일어났는지에 대한 검사비용을 줄이는 점에서 장점이 있다. 그러나 이러한 데이터 추출에 빈번하게 일어나는 공간데이터의 경우 갱신 질의가 발생할 때마다 동일한 공간데이터 추출에 따른 비용의 증가가 발생하는 단점이 존재한다. 그러므로 추가적으로 ODS에 임시 저장된 소스 시스템 별 갱신 질의에 대해 주제별로 통합된 데이터 추출이 필요하다. [그림 3]은 각 소스 시스템에서 발생된 갱신 질의에 대한

질의 통합 과정을 설명한다. 갱신질의에 대한 통합에서 처음 수행하는 갱신질의에 대한 수집단계는 각 소스 시스템으로부터 발생된 갱신 질의를 수집하고, 다음 단계로 각 시스템 별 갱신 질의에 대해 주제별 갱신 질의 분류를 수행한다.

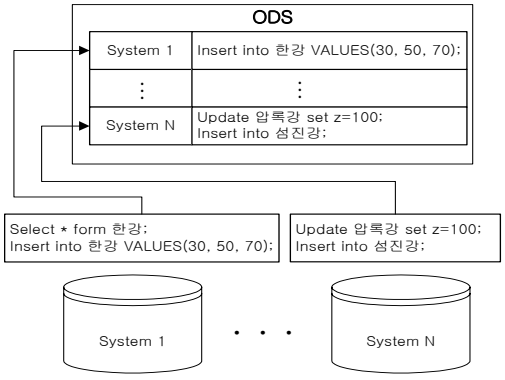


그림 2. ODS에 수집된 소스 시스템의 갱신질의 정보

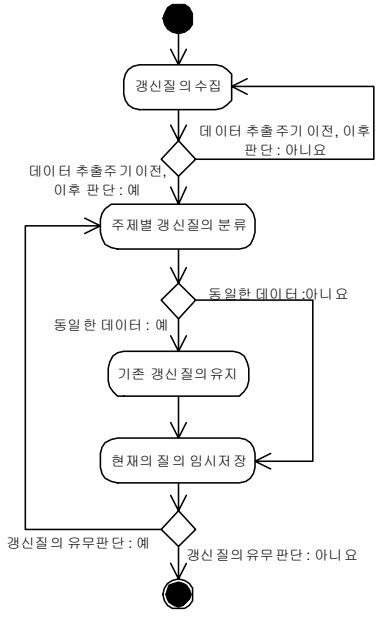


그림 3. ODS에 임시 저장된 갱신 질의에 대한 주제별 통합 과정

갱신 질의 분류단계에서는 관리자가 정의한 과거 추출 시점부터 현재의 추출시점 전까지 공간데이터에 갱신을 유발시킨 질의들에 대해 주제별 속성에 따라 분류를 진행한다. 즉, 동일한 주제의 데이터가 추출시점 이내에 다시 갱신이 발생했는지를 판단함으로써 같은 데이터에 반복적으로 발생하는 추출비용을 사전에 방지한다. 그러나 동일한 주제의 데이터에 대한 갱신질의가 아닐 경우 해당 질의를 유지함으로써 추출 주기가 이후 해당 질의에 대한 공간 데이터 추출을 수행한다.

이러한 ODS내에 임시 저장된 갱신질의에 대한 통합과정은 소스 시스템 별 남은 갱신 질의가 존재 하지 않을 때까지 수행한다.

3.2 이 기종 시스템에서의 공간 데이터 추출

소스 시스템으로부터 발생된 갱신 질의 추출 및 이에 대한 시스템 별 질의 통합 과정에 의해 생성된 주제별 통합 갱신 질의는 각 시스템에 추출을 위한 Select 질의로 변환되어 전송된다. 이렇게 시스템에게 재 전송되어 추출된 공간데이터는 기본적으로 이기종 시스템간 공간데이터 교환 표준인 GML 을 통해 얻어오며 시스템 별로 얻어온 공간데이터에 대해 주제별 중복 제거연산을 수행함으로써 SDW 시스템 내에 공간데이터의 중복 저장을 피할 수 있다.

ODS에서는 일반적으로 이 기종의 공간데이터 교환을 위한 표준 양식인 GML(Geography Markup Language)을 사용하여 이 기종의 공간데이터베이스 시스템으로부터 소스로 사용되는 다양한 공간데이터를 추출한다. 또한, GML로 추출한 공간데이터는 분석을 통하여 관계형 데이터로 변환 할 수 있다. [그림 4]는 이러한 GML 데이터에 대한 변환 예이다.

```
<?xml version="1.0" encoding="UTF-8"?>
<gml:name> Road </gml:name>
<gml:boundedBy>
<gml:Box srsName="http://www.opengis.net/gml/srs.epsg.xml#4326">
</gml:Box>
</gml:boundedBy>
<gml:LineString>
<gml:coord>
<gml:X> 0.000 </gml:X>
<gml:Y>          0.000 </gml:Y>
</gml:coord>
<gml:coord>
<gml:X> 3.000 </gml:X>
<gml:Y>          5.000 </gml:Y>
</gml:coord>
<gml:coord>
<gml:X> 5.000 </gml:X>
<gml:Y>          3.000 </gml:Y>
</gml:coord>
</gml:coord>
</gml:LineString>
```

(가) 이기종 소스로부터 추출된 GML 파일

< Road >

Type	Data
LineString	0.000,0.000 3.000,5.000 5.000,3.000

(나) GML 분석을 통해 추출된 관계형 데이터

그림 4. GML 데이터 분석을 통한 관계형 데이터 추출

일반적으로 GML 에서 관계형 데이터로의 변환 과정 중 이 기종 시스템에서 추출한 같은 주제의 공간데이터는 서로 다른 공간 타입으로 표현 될 수 있다. '한강' 이라는

이 기종에서 추출해 온 공간데이터가 있을 시, 소스 시스템 A에서는 해당 데이터를 PolyLine으로 표현할 수 있지만 소스 시스템 B에서는 Line으로 표현할 수 있다. 이와 같이 같은 데이터를 표현하는데 있어 이 기종이기 때문에 공간 타입의 불일치성이 존재한다. 이러한 문제를 해결하기 위해 시스템 관리자가 정의한 공간데이터 타입 정의 테이블을 참조하여 변환한다.

제안기법에서는 공간 데이터웨어하우스의 저장 비용을 줄이기 위해 소스 시스템으로부터 추출되는 공간데이터를 ODS의 현재 시점에 저장 중인 같은 주제의 공간데이터와 비교연산을 수행하여 변경된 공간 데이터만을 추출한다. 하지만 이러한 비교연산 시 현재 데이터와 과거데이터 사이에 변경이 발생하지 않는다면, 소스 시스템은 갱신정보 추출을 위해 과거에 적재된 과거 데이터와 비교연산을 반복적으로 수행해야 하는 단점이 있다. 따라서 변경된 데이터만을 추출하기 위해서는 공간 데이터웨어하우스에 기 적재된 데이터와 현재 시점에 ODS에 추출 및 통합된 공간데이터 사이의 비교연산이 필요하다. 이것은 각 시스템이 얻은 추출시점 별 갱신정보에 대하여 다음 추출시점에도 발생하는 중복된 갱신데이터의 추출을 제거할 수 있다. 또한, 공간 데이터간의 비중복 추출을 위한 비교연산을 이용하여 과거 시점과 현재 시점사이에 발생한 갱신 부분을 찾을 수 있다.

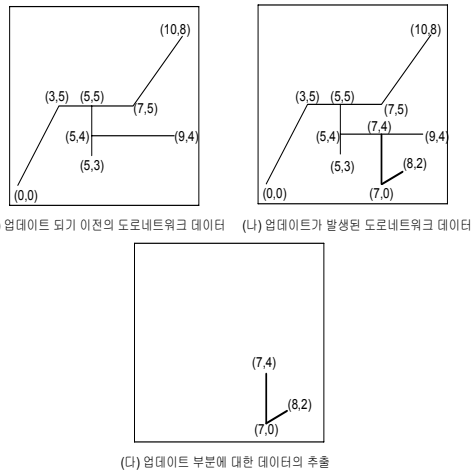
[알고리즘 1]은 현재 시점의 공간데이터와 공간 데이터웨어하우스의 과거 적재시점에 기 적재된 공간 데이터와의 비교연산을 이용하여 갱신된 공간데이터를 추출하기 위한 알고리즘이다.

[알고리즘 1] 갱신된 공간데이터 추출 알고리즘

```
Algorithm UDE(Current, Historical)
Input
Cur_Geom : 추출된 현재 시점의 공간데이터
Old_Geom : 과거 과거 시점의 공간데이터
Extract_Cnt : 추출 횟수
Output
Part_Geom: 추출된 부분 갱신 데이터
Err_Ext_Value : 잘못된 데이터추출
Begin
01 : for i:= 1 to i < Extract_Cnt-1 step 1
02 :   Part_Geom := Difference(Cur_Geom, Old_Geom)
03 :   if (Part_Geom != Null )
04 :     Old_Geom := Cur_Geom
05 :     Remove(Cur_Geom)
06 :   else
07 :     return ErrHandling(Err_Ext_Value)
08 :   end if
09 :   return Part_Geom
10 : end for
End
```

라인 1~2는 과거 공간데이터와의 비교연산을 수행함으로써 얻은 각 추출 시점마다의 갱신데이터 셋을 반환한다. 또한, 라인 3~5는 현재 시점의 공간데이터를 다음 추출시점의 비교연산에 사용될 통합된 공간데이터로 변환시킨다.

[그림 5]는 이러한 현재와 과거시점의 공간데이터에 대한 비교를 수행함으로써 현재 시점에서의 갱신된 공간데이터를 찾는 과정을 나타낸다.



(가) 업데이트 되기 이전의 도로네트워크 데이터 (나) 업데이트가 발생된 도로네트워크 데이터

(다) 업데이트 부분에 대한 데이터의 추출

그림 5. 도로 네트워크 공간 데이터에서 과거 데이터와의 비교를 통한 갱신 데이터 추출의 예

[그림 5]의 (가)는 공간 데이터웨어하우스에 저장되어 있는 갱신이 최종적으로 진행된 시점에 대한 공간데이터로 소스 시스템에서 갱신이 반영되어 추출된 (나) 와의 Difference 공간연산을 이용하여 (다)와 같이 변경된 공간데이터만을 추출한다. 이러한 공간 데이터웨어하우스 시스템 내부에 발생하는 중복된 데이터로 인한 저장 공간의 손실을 해결할 수 있다. 또한 중복 제거를 위한 공간연산은 공간 데이터웨어하우스에 대한 부하를 증가시키는 것이 아니라 전체 구성 시스템의 하부단계인 ODS, Wrapper 등의 구축기 단계에서 수행하는 것이므로 전체 공간 데이터웨어하우스 시스템에 대한 성능의 하향 및 부하가 적다.

4. 성능평가

본 장에서는 제안 기법인 비 중복 공간데이터 적재기법의 저장 공간 비용 및 데이터 적재횟수에 따른 비교와 데이터웨어하우스 시스템에서 가장 빈번하게 사용되는 집계 질의에 따른 응답시간을 제안기법과 기존기법에 대하여 실험을 통해 비교 평가한다.

실험은 Pentium(R)4 3.0GHz의 중앙처리장치, 2GB의 주 기억장치, 500GB 보조 기억장치의 IBM PC 호환기종에서 Windows XP Professional 환경에서 진행되었다.

본 실험에 사용된 평가 방법으로는 공간 데이터웨어하

우스 시스템으로 공간데이터 적재 시 적재시점 마다 추출된 데이터에 대해 일괄적인 적재방법(DSL: Duplicated Spatial data Loading)과 갱신된 데이터만을 추출하여 통합하는 본 제안기법과의 비교를 수행하였다. 제안되는 기법과 비교되는 DSL 적재방법은 비공간 데이터웨어하우스 시스템에서 수행하는 비공간 데이터에 대한 중복을 제거하지 않고 소스 시스템으로부터 추출한 데이터를 적재하는 방법으로 이것은 기존의 공간 데이터웨어하우스 구축을 위해 데이터 속성별로 공간데이터를 추출하여 레이어 단위 별로 시스템에 적재하는 방법과 동일하다. 즉, 제안하는 기법이 설명하는 데이터의 비 중복에 대한 성능평가 비교를 위해 기존의 일괄적인 데이터 적재기법과 비교하였다.

위 실험을 위한 소스 공간데이터로는 공간데이터 베이스 분야에서 많이 사용되는 TiGER/Line 파일을 사용하였으며, 공간 객체에 대한 적재횟수 및 응답시간을 기존 기법과 비교하였다[20]. 실험에 사용된 소스 공간데이터는 20,000개의 공간 레코드를 가진 20MB 크기의 임의로 생성된 데이터이다.

제안기법은 과거에 공간 데이터웨어하우스 시스템 내에 적재된 Historical 데이터와의 비교를 통해 현재 시점에 대한 갱신된 결과 데이터 셋 만을 적재한다. 이것은 기존 적재 기법이 데이터 속성별 레이어 단위를 기본으로 공간데이터를 적재하기 때문에 발생하는 저장 공간의 낭비를 해결할 수 있다. [그림 6]은 이러한 공간 데이터 적재 횟수에 따른 성능평가 측정 결과이다.

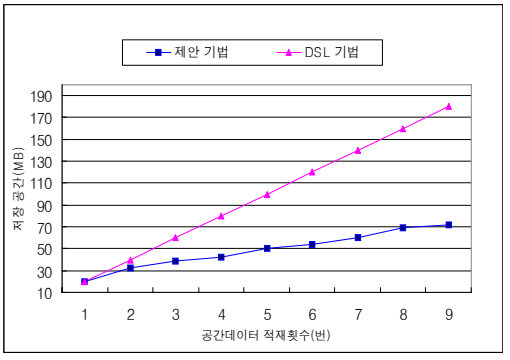


그림 6. 공간데이터의 적재 횟수에 따른 저장 공간의 크기 평가결과

기존의 적재기법은 매 공간데이터 적재 시마다 임의로 생성된 실험데이터 크기인 20MB의 크기씩 순차적으로 증가하는 모습을 볼 수 있다. 이것은 과거 시점에 적재된 데이터와의 비교를 수행하지 않기 때문에 나타난 결과로 이에 반하여 제안 기법은 매 적재 시점마다 갱신된 데이터 셋만을 저장하므로 기존 적재기법보다 적은 저장 공간 증가율을 보인다. 이러한 실험결과는 [그림 6]을 통해 기존 적재기법과 제안된 적재기법과의 저장 공간의 차이가 공간데이터의 적재횟수가 늘어날수록 커지는 것을 알 수 있다.

집계질의는 공간 데이터웨어하우스 시스템에서 강 유역의 변화, 도로 네트워크의 변화와 같은 일정 기간에 걸친 주제별 공간데이터의 변화에 대한 분석을 지원할 수 있다는 점에서 자주 사용되고 있다.

본 절에서는 이러한 데이터웨어하우스 시스템에서 대표적으로 사용되는 시간의 흐름에 따른 집계 질의에 대해 제안기법과 기존기법과의 질의 응답시간에 대한 성능평가를 수행한다. 제안기법은 시스템 적재 시점마다 소스 시스템으로부터 갱신된 데이터를 추출 및 통합 과정을 수행하여 최종적으로 단 하나의 갱신 데이터 셋만을 시스템 서버에 적재한다. 그러나 기존 적재기법에서는 주제별 하나의 레이어 단위로 각 소스 시스템에서 발생하는 갱신된 데이터를 시스템 서버에 적재하기 때문에 적재 시점마다 최악의 경우 소스 시스템 개수만큼 동일한 주제의 데이터가 존재 할 수 있다. [그림 7]은 이러한 집계질의에 따른 제안기법과 기존기법과의 응답시간 비교이다.

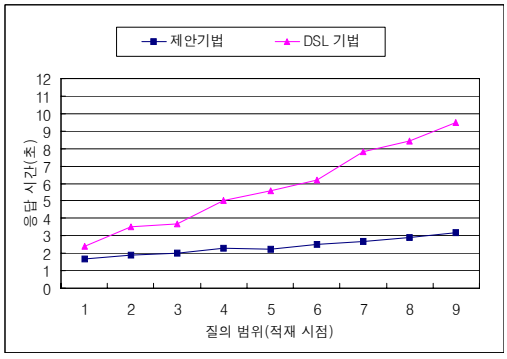


그림 7. 집계 질의 응답시간 평가결과

[그림 7]의 가로 축은 사용자의 질의에 참여하는 질의 범위이다. 즉, 사용자의 집계질의를 처리하는데 사용되는 시간 범위로 제안기법과 기존기법 모두 질의 범위가 늘어남에 따라 순차적으로 늘어나는 모습을 볼 수 있다.

그러나 제안기법은 시스템 적재 시점에 단 한 개의 통합 데이터를 가지고 있으므로, 기존 적재기법에 비해 질의 범위가 커지는 것에 대한 응답시간 상승폭이 작은 것을 볼 수 있다. 이것은 기존의 적재기법이 질의 범위내의 적재 시점마다 적어도 한 개 이상의 데이터를 가지고 있으며, 사용자의 집계질의 처리를 위해서는 적재 시점 별 해당 데이터들에 대한 사전 데이터 통합비용이 추가로 요구된다는 점 때문이다.

5. 결론 및 향후연구

본 논문에서는 공간 데이터웨어하우스에서 효율적인 공간 데이터 적재를 위한 이기종 간의 비중복 추출 기법을 제안하였다.

제안기법은 공간 데이터웨어하우스를 구축하기 위한 이 기종 소스로부터 공간데이터 추출 시 각 소스 시스템

으로부터 발생한 질의 중 데이터 갱신이 발생하는 삽입, 삭제, 업데이트에 대한 질의를 별도로 관리하고 해당 데이터만을 추출한다. 또한 공간 데이터웨어하우스 시스템에 기 적재된 같은 주제의 공간데이터와의 비교를 통해 갱신된 데이터만을 시스템내의 ODS에 유지하며 이러한 시스템 별 갱신된 데이터는 시스템 적재주기에 맞춰 주제별 하나의 업데이트 셋으로 공간 데이터웨어하우스 시스템에 적재된다. 제안 기법을 통해 주제별 데이터의 업데이트 셋만을 적재함으로써 기존 시스템이 가지고 있는 중복된 데이터의 저장으로 인한 공간 데이터웨어하우스 시스템 내부의 저장 공간의 비용을 줄일 수 있으며, 특정 시점에 대한 사용자 질의를 지원하기 위한 각 소스 시스템 별 추출된 주제별 데이터의 통합연산을 사전에 수행함으로써 해결하였다.

성능평가에서는 제안기법이 기존 제안기법에서 수행하는 주제별 레이어 단위의 적재방법과의 비교를 통해 적재 횟수가 증가할수록 저장 비용에서의 우수성을 보였으며, 사용자의 집계 데이터 검색에 대한 응답시간에 대해서는 약 50%의 성능 향상을 보였다. 그러나 초기 데이터 추출 후 적재에 대한 시스템 비용은 각 소스 시스템 별 갱신된 데이터의 통합연산을 수행하기 때문에 기존 적재기법에 비해 다소 낮은 성능을 보였다.

향후 연구로는 이러한 제안기법에서 수행하는 주제별 통합데이터 적재에 대한 과거 데이터와 소스로부터 추출된 데이터와의 비교연산 비용을 줄이는 기법에 대한 연구가 필요하다.

참 고 문 헌

[1] S. Chaudhuri, U. Dayal, “An Overview of Data Warehousing and OLAP Technology,” Proceedings of ACM International Conference on Management of data, ACM SIGMOD, Vol. 26, No. 1, 1997, pp. 65-74.

[2] W. H. Inmon, “Building the Data Warehouse,” 2nd Ed. John Wiley & Sons. Inc, 1996

[3] E. Sperley, “The EnterpriseData Warehouse: Planning, Building and Implementation,” Prentice Hall PTR, 1999, pp. 88-15.

[4] L. Savary, K. Zeitouni, “Spatial Data Warehouse - A Prototype,” A Proceedings of the EGOV2003, 2003, pp. 335-340.

[5] ESRI, “Spatial Data Warehousing for Hospital Organizations,” An ESRI White Paper, 1998. http://esri.com/library/whitepapers/pdfs/sdwho.pdf

[6] ESRI, “Spatial Data Warehousing,” An ESRI White Paper, 1998. http://www.geoweb.dnv.org/Education/whitepapers/SpatialWarehousing.pdf

[7] Oracle, “Oracle Spatial,” An Oracle White Paper, 2003. http://www.oracle.com/technology/products/spatial/pdf/spatial_best_practices.pdf

[8] 전병윤, 이동욱, 유병섭, 배해영, “공간 데이터웨어하우스에서 GML데이터의 효율적인 적재를 위한 데이터 통합기법,” 한국정보처리학회 2006년 춘계학술대회, Vol. 13, No. 1, 2006, pp. 27-30.

[9] 유병섭, 김경배, 이순조, 배해영, “공간 데이터 웨어하우스에서 공간 분석을 위한 공간 집계 연산,” 한국공간정보시스템학회 논문지, Vol. 9, No. 3, 2007, pp. 1-16.

[10] L. Stoimenov, S. Djordjevic, D. Stojanovic, “Integration of GIS Data Sources over the Internet Using Mediator and Wrapper Technology,” Proceedings of the 10th Mediterranean Electrotechnical Conference, Vol. 1, 2000, pp. 334-336.

[11] ESRI, “Spatial Data Standards and GIS Interoperability,” An ESRI White Paper, 2003. http://esri.com/library/whitepapers/pdfs/spatial-data-standards.pdf

[12] M. Howard, O. Dreza, “Combining Heterogeneous Spatial Data From Distributed Sources,” Proceedings of the 11thInternational Symposium on Spatial Data Handling, 2005, pp. 59-70.

[13] A. Simitsis, P. Vassiliadis, T. Sellis, “Optimizing ETL Process in Data Warehouse,” Proc. Of the 21st International Conference on Data Engineering, 2005, pp. 564-575.

[14] C. Squire, “Data extraction and transformation for the data warehouse,” A Proceedings of the ACM SIGMOD Internationalconference on Management of data, 1995, pp. 446-447.

[15] Oracle, “Integrated ETL and Modeling,” An Oracle White Paper, 2003. http://www.oracle.com/techno-logy/products/warehouse/pdf/OWB_WhitePaper.pdf

[16] 박동선, 배해영, “다차원 지리정보시스템을 위한 저장기법 및 분리된 저장구조,” 한국정보처리학회 논문지, Vol. 7, No. 1, 2000, pp 1-11.

[17] 전치수, 이동욱, 유병섭, 이순조, 배해영, “공간 데이터웨어하우스에서 시공간 분석 지원을 위한 비 중복 적재기법,” 한국공간정보시스템학회 논문지, Vol. 9, No. 2, 2007, pp. 81-91.

[18] D. Skoutas, A. Simitsis, “Designing ETL processes using semantic web technologies,” Proceedings Of the 9th ACM International workshop on Data warehousing and OLAP, 2006, pp. 67-74.

[19] X. CHEN, Z. CHI, X. CAO, “Applying DP to ETL of Spatial Data Warehouse,” 3rd Inter- national Conference on Machine Learning and Cybernetics, Vol.3, 2004, pp. 1616-1619.

[20] TIGER/Line Files, 2000 Technical Documentation, U.S. Bureau of Census, California, accessiblevia, http://arodate.esri.com/data/tiger2000/tiger_state-layer.cfm?stips=06

이 동 욱

2003년 상지대학교 전자계산공학과 (이학사)
2005년 인하대학교 컴퓨터 정보공학과 (공학석사)
2005년~현재 인하대학교 정보공학과 (박사과정)
관심분야 : 유비쿼터스 환경을 위한 공간 DBMS 및 DSMS, 공간 데이터웨어하우스

백 성 하

2005년 인하대학교 수학통계학부 (이학사)
2007년 인하대학교 컴퓨터 정보공학과 (공학석사)
2007년~현재 인하대학교 정보공학과 (박사과정)
관심분야 : 데이터 스트림 관리 시스템, 데이터베이스 클러스터

김 경 배

1992년 인하대학교 전자계산공학과 (공학사)
1994년 인하대학교 전자계산공학과 (공학석사)
2000년 인하대학교 전자계산공학과 (공학박사)
2004년~현재 서원대학교 컴퓨터교육학과 조교수
관심분야 : 이동실시간 데이터베이스, 스토리지 시스템, GIS

배 해 영

1974년 인하대학교 응용물리학과 (공학사)
1978년 연세대학교 전자계산학과 (공학석사)
1989년 숭실대학교 전자계산학과 (공학박사)
2006년~2009년 인하대학교 대학원 원장
1982년~현재 인하대학교 정보공학부 교수
관심분야: 공간 데이터베이스, 멀티미디어 데이터베이스 등