제 19 회 한국정보처리학회 춘계학술발표대회 논문집 제 10 권 제 1 호 (2003. 5)

# 시공간지원 집계 함수 설계*

신현호*, 최보윤*, 지정희*, 김상호*, 류근호*
*충북대학교 데이터베이스 연구실
e-mail : { shinhh, bychoi, jhchi, shkim, khryu @dblab.chungbuk.ac.kr}

# Design of Aggregate Function for Spatiotemporal

HyunHo Shin*, BoYoon Choi*, JeongHee Chi*, SangHo Kim*, KeunHo Ryu*
*Dept. of Computer Science, Chung-Buk National University

요 약

시공간 데이터베이스는 실세계에 존재하는 다양한 유형의 객체에 대한 공간 관리와 이력정보를 동시에 제공함으로써 사용자에게 시공간 데이터에 대한 저장 및 질의 수단을 제공한다. 질의 연산 중 집계 연산은 특정한 조건을 만족하는 데이터에 대하여 계산을 수행한 결과 값을 반환하는 연산으로, 다양한 분야에서 데이터의 분석을 위해 사용된다. 그러나 기존의 집계에 대한 연구는 시간 또는 공간에만 편중되어 시간과 공간 제약을 모두 가진 실세계의 응용에 직접 적용할 수 없다. 따라서 이 논문에서는 실세계 응용들의 분석을 위한 시공간 집계함수를 제안하고, 실제 응용에서의 분석을 위한 질의 예를 보인다. 제안된 시공간 집계함수에 의해 사용자는 응용시스템에 따른 시공간 데이터 분석을 위해 간략하고 편리한 질의 할 수 있다.

## 1. 서론

시공간 데이터베이스는 현실세계에 존재하고 있는 복잡하고 다양한 객체에 대하여 효율적인 공간정보를 제공할 뿐만 아니라 시간의 흐름에 따라 변화하는 이력을 동시에 효율적으로 관리함으로써 지리정보 시스템(geographic information system), 도시계획 시스템(urban plan system), 환경관리시스템(environment management system), 자동주행시스템(auto navigation system), 과학분야와 같은 광범위한 응용분야에서 사용될 수 있다[1].

지난 몇 년 동안 시간 및 공간 데이터베이스 분야에 대한 관심이 증가함에 여러 연구가 진행되었다. 초기에는 시공간 모델링, 색인기법, 저장구조에 관한 많은 연구가 수행되어 왔다[2,3,4,5]. 최근에는 기존의 데이터베이스 표준 질의 언어 SQL 에 시간 및 공간 개념을 추가한 질의언어에 대한 연구와 더불어 개념 추가로 인해 의미가 확장되는 함수 및 새롭게 정의 되는 함수와 이에 대한 처리기법에 관련된 연구가 수행되고 있다. 질의 함수 중에서 집계 연산(aggregation)은 릴레이션 전체 혹은 그 일부를 구성하는 튜플들에 적용되어 전체값을 계산하거나 대표값을 선택하는 연산으로 질의어에서 중요한 연산이다[6].

실세계 응용에서 ' 2001 년 8 월부터 2001 년 9 월까지 A 지역에 내린 강수량의 평균은 얼마인가?", " 2002 년 9 월 한달 동안의 쌀 수확량이 가장 많은 지역은 어느 지역인가?"와 같은 시공간 데이터에 대한 사용자 질의는 시간과 공간 제약을 모두 가지며, 과거 및 현재 상태에 대해 초점이 맞추어져 있다. 그러나 기존의 연구된 집계연산은 시간 또는 공간에 대해서만 제한된 집계 연산만 가능했다. 그러므로 시간과 공간의 특성을 모두 가지고 있는 실세계 응용에 직접 적용할 수 없다.

따라서, 이 논문에서는 시공간 데이터베이스와 시공간 모델을 바탕으로 시공간 데이터에 대한 공간관리와 이력관리의 효율적인 처리 및 과거와 현재상태에 대한 일련의 계산을 통해 부가적인 정보를 생성해 통계적인 분석을 할 수 있는 시공간 집계함수를 제안한다. 제안된 시공간 집계함수를 통해 질의 표현을 간략하게 할 수 있으며, 시간과 공간을 동시에 지원하는 질의 확장으로 사용자는 응용 시스템에 따른 시공간 데이터 분석을 위해 편리하게 질의를 할 수 있다.

효율적인 전개를 위해 이 논문의 구성은 2 장에서 시간 및 공간 집계함수에 대한 관련연구를 정리하고, 3 장에서는 시공간데이터를 위한 시공간 릴레이션을

* 이 논문은 2002 년도 한국학술진흥재단의 지원(KRF-2002-072-AM1013)에 의하여 연구되었음.

설계한다. 그리고 4 장에서는 시공간 집계함수 알고리즘을 설계하고 5 장에서 질의표현에 대해 기술하고 마지막으로 이 논문의 결론을 맺는다.

## 2. 관련연구

### 2.1 집계함수

집계함수(aggregate function)[7]는 데이터베이스에서의 기본적인 질의와는 달리 특정한 조건을 만족하는 데이터에 대하여 임의의 계산을 수행한 결과 값을 반환하며, 사용자로 하여금 질의 작성시 복잡한 질의를 간소화하고 시스템의 성능을 향상시키는 목적으로 사용된다.

관계형 데이터 모델에서 주로 사용하는 SQL 에서 지원하는 집계함수는 count, sum, min, max, avg 등이 있으며, Quel 에서는 count, sum, min, avg, any 등의 기본 집계함수와 유일한 값을 계산하기 위한 countU, sumU, avgU 등 유일 집계함수가 있다. 이러한 집계함수들은 min, max 등과 같이 특정 값을 찾아내는 선정 집계함수와 sum, avg 등과 같이 계산에 의하여 결과를 산출하는 계산 집계함수 및 count, any 등과 같이 개수를 계산하는 집계함수로 분류하며 이들 집계 함수들은 다음과 같이 두 가지 형태로 분류된다.

- 스칼라형(scalar)집계함수: 결과값으로 하나의 값을 생성한다.
- 관계형 집계함수: 대상 릴레이션의 한 부분 집합에 대해 주어진 집계함수를 계산하여 결정되는 여러 개의 값들을 생성한다.

### 2.2 시간 및 공간 집계함수

시간 지원 집계는 시간에 따라 변화하는 자료의 이력정보를 대상으로 수행된다. 이에 따라 일반적인 데이터베이스에서 시간차원을 추가한 새로운 집계함수에 대한 연구가 진행 되었다. 또한 시간지원 데이터베이스에서 집계 연산 처리를 위한 집계 트리(aggregation tree)기법[8]이 제안되었다. 이 기법은 기존의 집계 처리 연산 기법에서 확장된 시간 집계 연산 처리 기법들보다 우수한 성능을 나타낸다. 또한 디스크 저장 하기 위한 B+-tree 기반의 시간 집계 연산 처리 기법이 제안되었다[9,10].

공간 지원 집계는 질의를 위한 새로운 집계함수에 대한 연구가 진행되었다. 또한 효율적인 질의 처리를 위해 기존의 다차원 인덱스를 이용하는 MRA-tree(Multi-Resolution Aggregation tree)는 근접 집계 질의에 대한 처리를 할 수 있다[11]. 또한 공간 집계 함수 MIN, MAX 에 대한 집계 연산 처리 기법들이 연구 되었다[12].

그러나 시간 및 공간 지원 집계 연산 연구는 시간 또는 공간에만 편중된 정보만을 지원하므로 그 응용 분야에 따라 데이터 모델과 질의, 색인 구조 관련 많은 문제점 발생되고 있다. 따라서 이 논문에서는 제시된 문제점을 해결하기 위해 시공간 데이터에 대해 직접 적용할 수 있는 시공간 집계함수를 제안한다.

## 3. 시공간 데이터

### 3.1 시공간 데이터 모델

시공간 데이터베이스는 공간상에 존재하는 객체에 대하여 비 공간적인 요소뿐만 아니라 공간적인 요소, 시공간적 요소 그리고 이들에 대한 효율적인 관리와 서비스를 가능하게 해주는 새로운 데이터베이스 연구 분야이다. 시공간 데이터베이스는 공간객체의 공간정보 및 비공간 정보를 표현하고 공간에 따른 자료검색 및 갱신 등을 지원해야 하고, 시간에 따른 위치변화와 공간객체 상호간의 위상관계, 변화에 대한 이력 그리고 토지구역의 경우에는 시간의 흐름에 따라 소유주와 토지구역 형태의 변화까지 지원을 해야 한다. 이러한 기능을 지원하기 위하여 시공간데이터베이스는 시간의 흐름에 따라 변화되는 공간객체에 대한 이력자료를 모두 저장해야 한다.

이를 시공간 데이터 모델로 표현하기 위해 시공간 데이터베이스상에서 객체에 대한 효율적인 이력정보의 관리를 위해서 사용되는 두 가지 유형의 시간개념에는 거래시간과 유효시간이 있다. 거래시간이란 객체가 데이터베이스에서 처리된 시간으로 시스템에 의해서 자동적으로 부여되며, 유효시간은 실세계에서 객체가 유효한 시간으로 사용자에 의해서 결정된다[13]. 반면 객체에 대한 공간정보는 크게 래스터 유형과 벡터 유형으로 분류된다. 공간 연산에 사용되는 벡터 유형은 이미지 정보를 의미하는 래스터 유형의 정보로부터 유도되며 객체간 거리와 같은 기하 유형(geometric type)과 객체간 공간적 위치관계와 같은 위상 유형으로 세분화 한다[14].

이 논문에서는 위와 같은 시공간 데이터 모델을 기반으로 시공간 데이터 릴레이션 구조와 시공간 집계함수를 설계하였다.

### 3.2 시공간 데이터 릴레이션

이 연구에서는 공간속성의 시공간 특성을 고려한 데이터 저장 및 집계를 위해 관계형 시공간 데이터 모델을 적용한다. 관계형 시공간 데이터 모델은 일반적으로 임의의 객체에 대하여 내부적으로 시간 속성을 추가한 공간 테이블과 비공간 테이블로 구성하며 상호 조인을 통해 시공간 집계를 수행하는 특징을 가진다.

시공간 데이터베이스에서 표현 하기 위해 '공간 이력 릴레이션(spatial history relation)', '공간 속성 이력 릴레이션(spatial attribute history relation) ', '공간 정보 릴레이션(spatial data relation)'의 세 개의 릴레이션으로 구성된다. 공간 이력 릴레이션에는 시간의 변화에 따라 변경된 각 지역의 공간 정보가 저장되고, 공간 속성 이력 릴레이션은 시간의 변화에 따라 각 공간의 속성에 관련된 정보가 저장된다. 각 릴레이션에서 oid 는 릴레이션에서 객체 식별자이며, fid, rid 는 공간 또는 공간 속성 이력 릴레이션에서 공간 객체의 식별자(예: 객체의 주소), type 은 공간데이터 자료형( 예: 점, 선, 다각

형, 텍스트 ), MBR 은 공간 객체를 포함하는 최소경계 사각형의 최소 및 최대 좌표값, g_n 는 n 번째 공간의 속성 정보, VTs 는 유효시간 시작 시간, VTe 는 유효시간의 종료시간, TTs, TTe 는 트랜잭션 시간의 시작과 종료시간을 통해 이력정보를 표현한다.

표 1. 공간 이력 릴레이션 구조

| Oid | fid | Type | VTs | VTe | MBR | TTs | TTe |
|---|---|---|---|---|---|---|---|
| number | int | string | date | date | int | date | date |

표 1 의 공간 이력 릴레이션에서는 각각의 공간 객체에 대해 유효 시간 간격 별로 변화된 공간 정보가 저장된다.

표 2. 공간 속성 이력 릴레이션 구조

| Oid | rid | Type | VTs | VTe | g_1 | .. | g_n | TTs | TTe |
|---|---|---|---|---|---|---|---|---|---|
| number | int | string | date | date | float | .. | float | date | date |

표 2 공간 속성 이력 릴레이션에서는 각 공간의 유효 시간 간격 별로 변화된 속성 정보가 저장된다.

공간 이력 릴레이션과 공간 속성 이력 릴레이션에 저장되는 유효시간은 저장될 데이터의 특성에 따라 주기(granularity)를 서로 다르게 할 수 있다. 시간의 주기란 시간을 표현하는 최소 단위를 의미하며 년, 월, 일 등으로 구분할 수 있다. 단, 하나의 이력 릴레이션에서의 시간 주기는 모두 동일하다.

## 4. 시공간 집계함수

3 장에서의 시공간 데이터 모델과 릴레이션 구조를 기반으로 시공간 데이터베이스와 시공간 모델을 바탕으로 현실세계의 객체에 대하여 공간관리와 이력관리의 효율적인 처리 및 과거정보에 대한 일련의 계산을 통해 부가적인 정보를 생성해 통계적인 분석을 할 수 있는 시공간 집계함수를 설계하였다. 먼저 시공간 데이터베이스를 구성하는 지형 객체 fid 의 공간 정보와 공간 구성요소 그리고 기준객체와 탐색객체의 공통 유효시간을 정의 1, 2 에서 정의한다.

**정의 1.** 공간객체 fid 의 공간 속성은 공간 정보 S′(fid)와 공간 구성요소 SC′(fid)로 표현한다. Si(fid)는 객체의 fid 의 i 번째 공간 속성이라면 공간 정보 S′(fid)와 공간 구성요소 SC′(fid)는 다음과 같이 정의된다.

S′(fid)=[S1′(fid), S2′(fid), ....., Sn′(fid)] $F_{Li}$

SC′(fid)= [spatial components of a spatial object fid] $F_{Li}$

**정의 2.** 시간적으로 기준객체($W_{Li}$)와 탐색되는 객체($F_{Li}$)간의 공통된 유효시간(CommonValidtime)으로 정의한다.

정의된 공간요소와 CommonValidtime 으로 이 논문에서 제안한 시공간 집계함수의 정의는 다음과 같다.

**정의 3.** stCOUNT 집계함수는 주어진 시공간 객체 $W_{Li}$ 의 유효시간에 포함되고 공간적으로 시공간 객체 $W_{Li}$ 에 포함되는 시공간 객체 $F_{Li}$ 의 모든 개수를 나타낸다. 그 관계식은 다음과 같이 표현된다.

Spatial($F_{Li}$) stCOUNT Spatial($W_{Li}$)
= [COUNT(∀ $F_{Li}$ ) with ∃SC′($F_{Li}$) ∩ ∃SC′($W_{Li}$)
∧ CommonValidtime($F_{Li}$, $W_{Li}$)]

**정의 4.** stSUM 집계함수는 주어진 시공간 객체 $W_{Li}$ 의 유효시간에 포함되고 공간적으로 시공간 객체 $W_{Li}$ 에 포함되는 시공간 객체 $F_{Li}$ 들의 속성 값의 합을 나타낸다. 그 관계식은 다음과 같이 표현된다.

Spatial($F_{Li}$) stSUM Spatial($W_{Li}$)
= [SUM(∀A′(fid)) with ∃SC′($F_{Li}$) ∩ ∃SC′($W_{Li}$)
∧ CommonValidtime($F_{Li}$, $W_{Li}$)]

**정의 5.** stAVG 집계함수는 주어진 시공간 객체 $W_{Li}$ 의 유효시간에 포함되고 공간적으로 시공간 객체 $W_{Li}$ 에 포함되는 시공간 객체 $F_{Li}$ 들의 stSUM/stCOUNT 된 속성 값이다. 그 관계식은 다음과 같이 표현된다.

Spatial($F_{Li}$) stAVG Spatial($W_{Li}$)
=[AVG(SUM(A′(fid )) /COUNT ($F_{Li}$))
with ∃SC′($F_{Li}$)∩∃SC′($W_{Li}$)∧CommonValidtime($F_{Li}$, $W_{Li}$)]

**정의 6.** stMAX 집계함수는 주어진 시공간 객체 $W_{Li}$ 의 유효시간에 포함되고 공간적으로 시공간 객체 $W_{Li}$ 에 포함되는 시공간 객체 $F_{Li}$ 들의 속성들 중 가장 큰 속성값의 객체 $F_{Li}$ 이다. 그 관계식은 다음과 같이 표현된다.

Spatial($F_{Li}$) stMAX Spatial($W_{Li}$)
= [MAX($F_{Li}$) with ∃SC′($F_{Li}$) ∩ ∃SC′($W_{Li}$)
∧ CommonValidtime($F_{Li}$, $W_{Li}$)]

**정의 7.** stMIN ⇔ stMAX

정의된 시공간 집계함수들 중에서 시공간 객체의 속성값의 합을 반환하는 stSUM 알고리즘은 표 3 과 같다.

표 3. stSUM 집계함수 알고리즘

```
Spatiotemporal Aggregate Function stSUM
Input : feature fid, spatial object type
Output : sum of feature attribute
While(현재 feature 스키마가 존재함) {
  If (현재 feature 스키마 내의 공간 식별자와 입력된 튜
  플의 공간 식별자가 일치함) {
      if(현재 feature 스키마 내의 튜플이 가지는 유효기간
        과 입력된 튜플이 가지는 유효기간이 일치함) {
          for() {
              반환값+= 집계된 속성값;
              동일 튜플 내의 다음 집계 속성으로 이동;
                  }
              }
          }
      다음 튜플 검색;
}
```

이 논문에서는 시공간 데이터를 지원하는 stCOUNT, stSUM, stAVG, stMAX, stMIN 집계 함수를 제안하였다. 각각 stCOUNT, stMAX, stMIN 은 주어진 유효시간 동안 만족하는 특정 값을 찾아내는 집계 연산이며,

stSUM, stAVG 는 주어진 유효시간 동안 계산에 의해 결과를 산출하는 집계 연산이다. 시공간 데이터의 공간객체와 유효시간에 대한 정보를 비교 함으로써 집계 연산을 수행한다.

## 5. 시공간 집계 질의 예

설계된 시공간 집계함수를 기반으로 사용자 질의를 표현 해보면 다음과 같다. 질의 예는 농경지와 강수지역의 릴레이션을 사용하였다.

**[질의 1]** " 농경지에서 1999 년 6 월부터 8 월까지 A 지역에 뿌려진 농약의 총량은 얼마인가?"

```
SQL > SELECT   A. stSUM
    > FROM   Field f
    > WHERE   fid= A and
        f.valid overlap PERIOD '01-Jun-99, 01-Aug-99';
```

위의 질의는 stSUM 시공간 집계 함수를 통해서 얻어진다. Field 의 '공간 이력 릴레이션' 으로부터 식별자 'A'와 유효기간 'Jun-01-99, Aug-01-99'를 가진 레코드를 검색한다. 이 레코드들의 속성정보는 '공간 속성 릴레이션' 으로부터 주어진 유효시간과 부합되는 한 반복적으로 탐색한다. 나머지는 기존의 함수를 통해서 사용자 질의를 처리해 준다.

**[질의 2]** " 서울에서 2000 년 6 월부터 9 월까지 T 지역의 평균 강수량은 얼마인가?"

```
SQL > SELECT   T.stAVG
    > FROM   Seoul s
    > WHERE   fid= T and
        s.valid overlap PERIOD '01-Jun-00, 01-Sep-00';
```

주어진 영역과 기간 내에 강수량의 평균은 stAVG 함수를 이용하여 얻어진다.

이 논문에서 제안한 시공간 집계함수에 의한 질의 표현은 질의 1,2 와 같이 질의가 간략하게 표현되며, 질의 표현의 편리성과 간결함을 제공하여 사용자는 응용시스템에 따른 시공간 데이터에 대한 질의를 쉽게 할 수 있다.

## 6. 결론

실세계의 객체는 시간에 따라 변화하는 이력정보와 객체의 공간정보를 가지고 있다. 이에 따라 시공간 객체의 이력관리와 공간정보의 효율적인 처리 및 과거 정보에 대한 통계적인 분석이 대두되었다. 그러나, 기존의 집계함수에 대한 연구는 시간지원과 공간지원에 대해 각각 독립적인 연구만이 수행되어 시공간 데이터에 직접적으로 적용 할 수 없는 문제가 있다.

이와 같은 문제를 해결하기 위해 이 논문에서는 시공간 객체를 표현하기 위한 데이터베이스의 구조를 제시하고 이를 바탕으로 시공간 데이터베이스에 저장된 다양한 유형의 시공간 객체에 대한 이력정보와 공간정보의 분석이 가능한 시공간 데이터 지원 집계함수 stCOUNT, stSUM, stAVG, stMAX, stMIN 를 제안하였다. stCOUNT, stMAX, stMIN 은 주어진 유효시간 동안 만족하는 특정 값을 찾아내는 집계 연산이며, stSUM, stAVG 는 주어진 유효시간 동안 계산에 의해 결과를 산출하는 집계 연산이다. 또한 시공간 집계함수를 사용한 질의 예제를 통해 기존의 시간, 공간 집계함수가 분석할 수 없었던 시공간 응용에 대한 분석의 용이함을 보였다.

최근 들어 많은 응용 시스템은 그 분야에 따라 관계 질의와 함께 공간 질의, 시간질의, 시공간 질의의 확장을 요구하고 있다. 따라서, 이 논문에서 제안된 시공간 집계함수를 통해 질의 표현을 간략하게 할 수 있으며, 시간과 공간을 동시에 지원하는 질의 확장으로 사용자의 응용 시스템에 맞는 질의 표현력의 개선 및 편리성을 제공한다. 또한 시공간 질의어 상에서 불필요한 연산의 사용을 방지 함으로써 전반적인 시스템 성능의 향상을 기대할 수 있다.

향후 연구로는 상세한 분석이 가능한 확장된 시공간 집계함수를 제안하고 제안한 집계함수를 구현하는 작업이 남아 있다. 또한 실제 응용 시스템에 이 집계 함수를 적용 및 평가하는 작업이 수행될 것이다.

## 참고문헌


[1] M. Worboys, "A Unified Model for Spatial and Temporal Information," The Computer Journal, Vol.37, No.1, 1994.

[2] Volker Gaede and Oliver Gunter. "Multidimensional Access Methods," ACM Computing Surveys, 1997.

[3] Open GIS Consortium, Inc. OpenGIS, "Simple Features Specification For OLE/COM Revision 1.1," OpenGIS Project Document, 99-050, 1999.

[4] M.Nascimento and J.O.Silva, "TOWARDS HISTORICAL R-TREES,"ACM, pp.235-240, 1998.

[5] T.K.Sellis, "Research Issues in Spatio-temporal Database System," SSD, pp.5-11, 1999.

[6] J.Gray, "The Benchmark Handbook for Database and Transaction Processing Systems," Morgan Kaumann, 1991.

[7] 김동호, 이인홍, 류근호, "주기억장치에서 시간지원 데이터 베이스의 집계함수 설계 및 구현," 정보과학회 논문지, 제 21 권, 제 8 호, 1994 년 8 월.

[8] N.Kline, and R.T.Snodgrass, "Computing Temporal Aggregates,"International Conference on Database Engineering, p222-231, 1995.

[9] D. Zhang, A. Markowetz, V. Tsotras, D. Gunopulos and .Seeger, "Efficient Computation of Temporal Aggregates with Range Predicates," in PODS 01, Santa Barbara, CA, 2001.

[10] J.Yang and J.Widom, "Incremental Computation and Maintenance of Temporal Aggregtes," International Conference on Data Engineering, pp.51-60, 2001.

[11] I.Lazaridis and S.Mehrotra. "Progressive approximate aggregate queries with a multi-resolution tree structure," SIGMOD, 2001.

[12] Donhui Zhang and Vassilis J.Tsotras, "Improving Min/Max Aggregation over Spatial Object," ACM-GIS, 2001.

[13] R. Snodgrass, "The Temporal Query Language TQuel," ACM TODS. Vol.12, No.2, Jun, 1987.

[14] P.Svensson and Z, Huang, "A Query Language for Spatial Data Analysis," Advances in Spatial Databases, SSD 91, Aug, 1991.