u-GIS 환경에서 다중 공간 집계 질의의 중복연산 비용을 감소시키기 위한 자원공유 기법

서민호*, 김상기*, 백성하*, 이연*, 이동욱*, 배해영*
*인하대학교 컴퓨터 정보공학과
e-mail : {mhseo, kimsk, shbaek, leeyeon, dwlee }@dblab.inha.ac.kr, hybae@inha.ac.kr

Resource Sharing Method to Reduce Duplicate Operation Cost of Multiple Spatial Aggregates in u-GIS Environment

Min-ho Seo *, Sang-Ki Kim*, Sung-Ha Baek*, Yan Li*, Dong-Wook Lee*, Hae-Young Bae*
*Dept. of Computer Science, In-ha University

요 약

데이터 스트림을 처리하기 위한 연속집계질의 수행 시 중복연산 및 메모리의 절약을 위하여 큐를 공유하는 자원공유기법이 연구되었다. 기존의 자원공유 기법들은 질의의 프리디킷이 일치할 때만 처리하기 때문에, 질의의 프리디킷이 차이가 나는 경우가 많은 다중공간 집계질의가 자주 요청되는 u-GIS 환경에서 효율적으로 중복영역을 처리할 수 있는 자원공유 기법이 요구된다. 본 논문에서는 공간영역을 효율적으로 그룹화하는 R-tree 의 특징을 이용하여 질의간의 중복영역을 그룹화하고 중복영역의 자원을 패인(Pane)구조를 이용하여 공유한다. 노드 수에 제한이 없고 레벨을 1 로 하는 R-tree 로 유사한 위치의 질의들을 그룹화 한 후, 그 질의들의 영역이 겹쳐지는 부분을 패인을 이용해 집계 값을 공유하여 중복계산을 피하는 방법이다. 제안 기법은 공간 집계질의를 처리할 수 있고, 기존의 계층구조의 자원공유 기법을 사용할 때에 비해 자원을 적게 사용하고 질의 처리 시간을 단축시켰다. 성능평가를 통하여 제안기법이 메모리 사용량을 감소시키는 것을 보였으며, 질의 처리 속도가 증가하였다.

1. 서론

최근 센서 네트워크 및 통신기술의 발달에 의해 데이터 스트림(Data Stream)에 대한 연구가 활발히 진행중이다[1]. 특히, 다양한 종류의 컴퓨터가 사람, 사물, 환경 속으로 스며들고, 이들이 서로 네트워크로 연결되는 유비쿼터스 환경의 도래와 더불어 공간 및 위치정보를 제공하는 공간정보 기술이 급속히 발전하고 있다. 기존의 GIS 는 건물, 도로, 하천과 같은 2 차원 또는 3 차원상의 정적인 지형지물 관리에만 초점을 맞추어 왔지만, u-GIS 에서는 유비쿼터스 환경을 기반으로 시간에 따라 공간적인 위치가 포함된 동적인 공간정보의 활용으로 영역을 확장하고 있다[2,3]. 또한, GeoSensor 에서 수집되는 데이터 스트림과 GIS 데이터, 그리고 공간질의를 융합하여 처리하는 u-GIS 공간기술이 연구되고 있다.

u-GIS 에서는 GeoSensor 로부터 유입되는 위치를 포함한 데이터 스트림을 처리하기 위해 공간 연속질의(Continuous Query Language)[1,4]를 사용한다. 데이터 스트림이 매우 빠르게 발생하고 시스템에 다수의 질의가 등록되면 질의들이 소유하고 있는 독립적인 큐를 유지하기 위한 메모리 비용이 증가하고 동일한 연산을 반복하는 문제가 발생한다. 이 비용들을 줄이고자 집계질의 처리 시 큐를 공유하는 자원공유 기법이 연구되었다[5].

DSMS 에서 자원을 공유하기 위한 기법으로 BINT, LINT 그리고 패인기법이 제안되었다. BINT 는 입력된 튜플들을 최하 단부터 계산하여 집계정보를 저장해 놓는 방법으로, 질의 별로 큐를 유지하지 않는 장점이 있지만, 계층별 집계정보의 저장으로 인해 메모리 유지비용 증가와 집계정보 검색 시간이 증가한다. LINT 는 집계정보를 좌우 대칭으로 저장하여 질의범위의 검색시간을 감소시켰지만, 좌우 대칭구조를 사용하여 메모리 비용이 BINT 보다 많이 든다. 마지막으로 패인(Pane)구조를 이용한 방법은 질의들의 질의범위의 최대공약수 크기만큼 집계정보를 저장하여 동일한 튜플이 등록된 질의에 의해 반복적으로 계산되는 것을 방지해 집계정보 생성시간을 단축시켰다.

이러한 기존의 자원공유 기법들은 비공간 데이터스트림의 처리만을 고려한 방법들이다. 기존의 자원 공유 기법을 공간 집계 질의에 적용할 경우, 각 질의의 모든 영역에 대해 BINT 를 구성해야 하는 문제가 발생한다. 방대한 데이터 양을 다루는 공간 집계 질의의 특성 때문에, BINT 를 사용한 방법은 메모리 유지비용과 질의처리시간이 너무 많이 소요된다. 또한, 패인을 적용하였을 경우, 패인의 집계 값의 저장단위인

본 연구는 건설교통부 첨단도시기술개발사업 - 지능형국토정보기술혁신 사업과제의 연구비지원(07 국토정보 C05)에 의해 수행되었습니다

질의범위와 갱신범위의 최대공약수(GCD)가 1 이 되는 경우에는 튜플을 하나씩 계산하는 경우와 질의처리 효율이 같게 된다. 그러므로, 다중 공간 집계질의가 자주 요청되는 u-GIS 환경에서 중복된 영역을 최대한으로 공유하여 질의를 처리할 수 있는 자원공유기법이 요구된다.

본 논문에서는 공간영역을 효율적으로 그룹화하는 R-tree 의 특징을 이용하여 질의간의 중복영역을 찾고, 중복영역의 자원을 공유할 수 있는 확장된 패인 기법을 제안한다. 기존 R-tree 의 한 노드 당 저장할 수 있는 데이터 수의 제한은 공간적으로 거의 유사한 영역을 분할할 수 밖에 없게 한다. 그러한 R-tree 의 제약을 완화하여 각 노드가 갖는 데이터 수의 제한을 없애고, 영역 확장을 최소화하는 분할 기법인 Quadratic Split 을 이용하여 그룹화를 함으로써 질의간의 최대 공유부분을 알 수 있다. 그리고, 그룹내의 중복영역들에 패인을 적용하여 자원을 공유한다. 중복되는 질의영역에 대해서만 패인을 구성하여 자원을 공유함으로써 모든 질의영역에 대해 BINT 를 적용해야 하는 문제가 해결되고, 패인의 단위를 구성하는 GCD 가 1 이 되는 확률이 줄어든다. 결과적으로, 질의처리 속도가 향상되고 메모리의 사용량도 줄어든다.

본 논문의 구성은 다음과 같다. 2 장에서는 u-GIS 공간 정보 기술과 기존의 자원공유 기법들에 대해 알아보고, 3 장에서는 본 논문에서 제안한 다중 공간 집계질의의 중복을 효율적으로 처리하기 위한 자원공유 기법에 대해 설명한다. 4 장에서는 제안한 기법의 성능 측정 결과를 평가하고, 마지막으로 5 장에서는 본 논문의 결론 및 향후 연구를 보인다.

2. 관련연구

본 장에서는 본 논문의 기반이 되는 u-GIS 공간 정보 기술과 자원공유를 이용한 연속 집계질의 처리 기법인 BINT, LINT 그리고 패인구조에 대해 기술한다.

2.1 u-GIS 공간 정보 기술

u-GIS 공간정보 기술은 기존 GIS 인 건물, 도로, 하천, 지하시설물과 같은 2 차원 또는 3 차원상의 정적인 지형 지물 정보와 유비쿼터스 환경을 기반으로 시간에 따라 공간적인 위치가 포함된 동적인 GeoSensor 정보의 융합 처리를 요구한다. GeoSensor 는 고정된 지역이 아니라 넓은 지역에 산발적으로 분포될 수 있으며, 자신의 위치인식 장치를 이용하여 시간에 따라 장소를 이동할 수 있는 이동성도 가지고 있다. 따라서 GeoSensor 는 데이터 스트림의 특성과 GIS 의 특성을 동시에 갖는다. GeoSensor 는 넓은 지역에서 실시간으로 발생하는 대용량 정보를 처리하기 위해 데이터 스트림 처리 기술이 요구된다. 또한 GIS 공간 정보와의 융합 처리를 위해 공간 연산 처리가 요구된다.[2,3]

2.2 자원공유기법

스트림 환경에서 다중 집계질의 처리를 위한 자원공유 기법으로 BINT(Base-Interval), LINT(Landmark-Interval) 그리고 패인(Pane)기법이 있다[6,7]. BINT 는 범위(Interval)와 계층(level)을 사용한다. 범위는 집계정보를 갖는 단위를 의미하며 계층은 각 계층마다 포함하고 있는 집계 값의 튜플 개수를 의미한다. 집계정보 생성은 튜플이 입력될 때 마다 발생하는데, 먼저 입력된 튜플들을 최하위 계층에 저장한다. 그리고 입력된 두 개의 튜플들로 집계정보를 계산하고 그것을 상위 계층에 저장한다. 이렇게 상위 계층에 생성된 집계정보 두 개를 가지고 하나의 집계정보를 계산하여 다시 상위 계층에 저장한다. 이런 방법으로 집계정보를 점차 상위계층으로 저장하여 계층구조의 자료구조를 생성한다. 자료구조 생성 후, BINT 는 질의처리를 위하여 질의 범위를 만족하는 집계정보를 검색하고, 각 집계정보에 저장된 집계 값들을 계산하여 질의 결과를 반환한다. BINT 는 질의 처리 시 집계 값을 중복 계산 하지 않는 장점이 있지만, 튜플을 기본 단위로 집계정보를 저장해야 하기 때문에 집계정보를 생성하는 시간이 많이 필요하고 질의처리 시간 또한 늘어난다.

유사한 기법인 LINT 는 집계정보를 저장한 계층구조를 좌우로 구축하여 두 개의 영역만 검색하면 질의 결과를 얻을 수 있다. 그래서 BINT 의 단점인 질의 범위를 만족하는 범위를 검색하는 속도를 향상시켰다. 그러나 두 개의 계층구조를 사용하기 때문에 계층구조 구축시간과 공간 낭비가 BINT 보다 크다.

마지막으로 패인(Pane)구조가 있다. 슬라이딩 윈도우는 질의범위가 갱신범위보다 크기 때문에 질의처리를 위해 스트림의 버퍼를 이동할 때 중복된 영역을 포함하게 된다. 이 중복된 영역을 반복 계산 하지 않기 위해 패인구조가 제안되었다. 패인은 갱신범위와 질의범위의 최대공약수를 계산하여 그것을 단위로 입력 데이터들을 분할하고, 분할 된 튜플들의 집계 값을 저장한다. 질의 처리 시에 분할 저장된 패인의 집계 값을 이용하여 질의 결과를 반환하기 때문에 튜플에서 집계 값을 구하는 연산은 패인을 생성할 때 한 번만 수행된다.

위와 같은 기존의 자원공유 기법들은 비공간 데이터 스트림을 다루는 연속질의 처리만을 고려한 방법이다. 기존의 자원공유 기법으로는 공간 질의를 처리할 때 질의영역이 겹쳐지는 부분에 대한 자원공유를 할 수가 없다. 또한, 방대한 데이터를 다루는 공간 질의의 특성 때문에, 모든 튜플들의 집계정보를 계층적으로 저장하는 BINT 는 메모리의 비용이 많이 소요되기 때문에 공간질의의 자원공유 기법으로 부적합하다. 공간정보를 지닌 공간질의에서 중복된 영역을 최소화하고, 중복영역의 자원을 효율적으로 공유할 수 있는 방법이 필요하다.

3. 공간 집계질의에서의 자원공유 기법

본 장에서는 공간영역을 효율적으로 그룹화하는 R-tree[8]의 특징을 이용하여 질의간의 중복영역을 찾고, 중복영역의 자원을 공유할 수 있는 확장된 패인기법에 대해서 설명한다. 제안기법은 다중 공간 집계질의의 중복영역을 확장된 R-tree 를 이용해 그룹화 한 후, 그룹화 한 영역에서 질의가 겹쳐지는 영역들에 대해

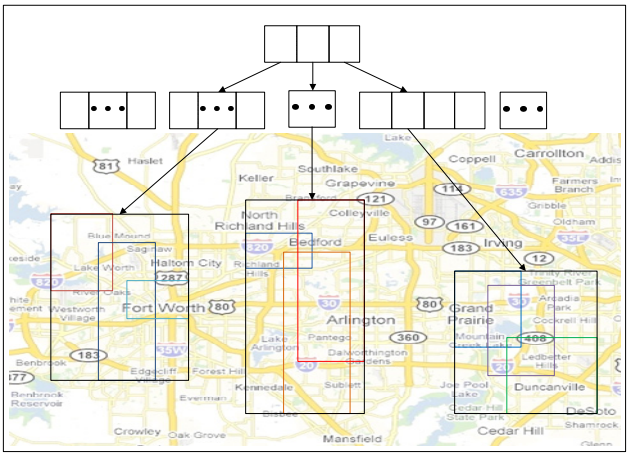
패인을 구성하여 자원을 공유하는 기법이다. 본 장 3.1 절에서는 확장된 R-tree 를 이용한 다중 공간 집계 질의의 그룹화에 대해서 설명하고, 3.2 절에서는 그룹화한 영역에서 공간 집계질의들이 중복되는 영역에 대해 패인을 이용해 자원을 공유하는 방법을 보인다.

3.1 omR-tree 를 이용한 그룹화

R-tree 는 B-tree 와 유사한 공간 인덱스 기법이다. R-tree 의 장점은 각 공간을 MBR 로 구성, 분할하여 그룹화함으로써 노드의 중복을 최소화 할 수 있다. 또한, 노드의 최대 차수는 3~4 레벨을 유지하며 좌우노드의 균형을 유지하기 때문에 검색시간이 빠른 특징을 갖는다.

본 논문에서는 다중 공간 집계질의들의 중복된 영역을 찾아내기 위해 omR-tree(One Level Multi node R-tree)를 사용한다. omR-tree 는 R-tree 를 기반으로 질의의 중복 영역을 최소화하기 위해 R-tree 의 Quadratic Split 알고리즘을 사용한다. 또한, 데이터의 빠른 삽입을 위해 omR-tree 높이를 1 레벨로 제한하고 중복된 영역의 더 많은 자원을 공유하기 위해 각 노드가 저장할 수 있는 데이터의 양에 제한을 두지 않는다.

다음의 (그림 1)은 omR-tree 를 이용해 다중 공간집계질의의 그룹화의 한 예를 보여준다.

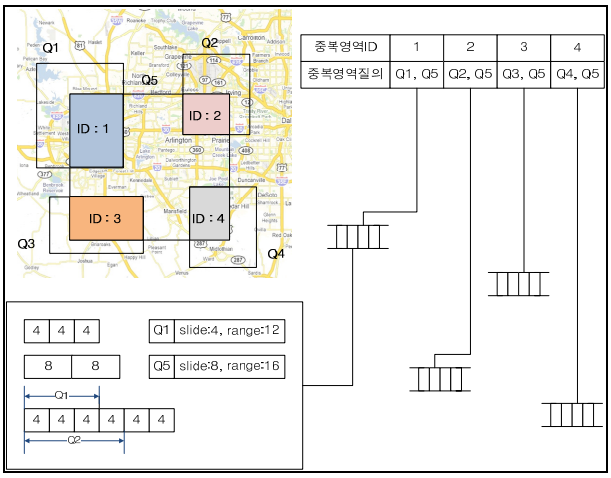


(그림 1) omR-tree 를 이용한 다중 공간집계질의의 그룹화의 예

3.2 패인을 이용한 공유방법

omR-tree 를 이용하여 그룹화 된 노드의 중복된 영역에 대한 공유된 자원을 유지하기 위해 패인기반의 자원 공유 기법을 사용한다. 우선 그룹화 된 한 영역에서 각 질의들이 겹쳐지는 영역을 식별한다. 질의가 중복되는 각 영역에 ID 를 부여하고, 각 중복된 영역마다 메모리 큐를 할당한다. 중복되는 영역이 D_1~D_n 이라 할 때, 각 중복된 영역마다 메모리 큐를 Q_1~Q_n 까지 할당한다. 그리고 중복된 영역의 집계질의 정보와 중복영역의 ID 를 포함하는 테이블로 큐의 정보를 관리한다.

할당된 메모리 큐는 패인기반의 S-패인(Spatial Pane)을 적용한다. S-패인은 질의들의 갱신범위(Slide)와 질의범위(Range)들의 최대공약수를 단위로 입력데이터를 분할하고, 분할 된 튜플들의 집계 값을 저장한다. S-패인은 GCDs-pane(GCDq$_i$(Rangeq_i, Slideq_i), … GCDq$_j$(Rangeq_j, Slideq_j))으로 계산하여 데이터 분할 단위를 결정한다. 데이터는 유입되면서 해당 공간영역에 따라 각 큐에 쌓이게 되고 S-패인을 이용해 미리 집계 값을 계산한다. 질의 처리시에는 분할 저장된 패인의 집계 값을 공유하여 사용하기 때문에 중복 계산을 감소시켜 질의 처리 속도를 향상시킨다.

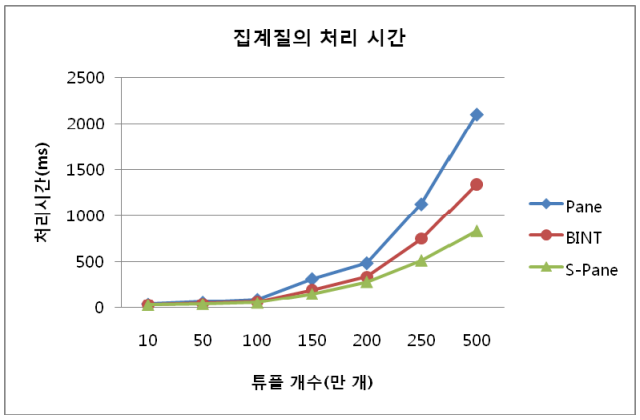


(그림 2) 그룹화된 영역에서 S-패인을 이용해 중복영역에 대해 자원을 공유하는 예

(그림 2)는 omR-tree 를 이용하여 공간집계 질의들을 그룹화 한 후, 그룹화된 영역에서 질의들의 중복영역을 S-패인을 이용하여 공유하는 예를 보여주고 있다. 우선 질의가 겹치는 부분들에 대해서 ID 를 부여하고 질의들의 정보를 테이블에 저장한다. 그리고 중복되는 영역마다 메모리 큐를 할당한다. 그림에서는 네 곳의 중복영역이 있으므로 큐를 4 개 할당하게 된다. 그 다음 큐를 S-pane 구조로 만들기 위해 데이터를 분할할 단위를 만든다. (그림 2)에서와 같이 질의 1 과 질의 5 가 겹쳐지는 부분에 대해 S-패인을 구성하는 법은 다음과 같다. GCDs-pane(GCDQ1(12, 4), GCD(Q5(16, 8)) = 4 이므로, S-패인의 데이터 분할 단위는 4 가 된다.

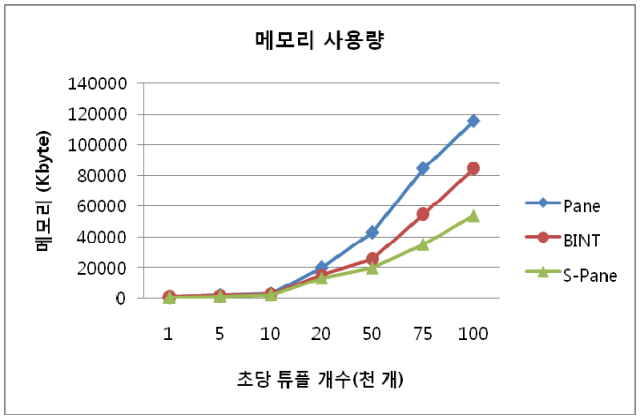
4. 성능평가

성능평가는 공간 정보를 포함하는 데이터 스트림을 처리하는 환경에서 공간 집계질의를 그냥 처리했을 경우와 공유되는 부분을 BINT 로 구성해 처리했을 때, 그리고 제안기법을 사용할 때를 서로 비교 분석하였다. 모든 실험은 2GB 메모리와 Intel Xeon 1.86GHz CPU 를 가진 HP ProLiant DL380 G5 서버에서 Window 2003 Server 운영체제 하에서 수행되었으며 알고리즘 구현은 Microsoft Visual Studio 2005 에서 C++로 구현하였다. 그리고 실험에 사용된 데이터는 250,000 개 이고 튜플 스키마는 <ts, area, temperature> 이며, ts 는 8byte, area 16byte, temperature 4byte 로 총 28byte 의 크기를 갖는다.



(그림 3) 주어진 튜플에서 집계질의 처리 시간

(그림 3)은 주어진 튜플에서 집계질의를 처리하는 시간의 비교이다. Normal 은 튜플 자체를 공유하기 때문에 슬라이딩 윈도우가 이동하면서, 동일한 튜플을 재계산하게 되므로 처리속도가 상당히 느리다. 그리고 BINT 는 계층구조로 집계 정보를 저장하여, 질의 수행영역을 만족하는 집계 정보를 찾는 비용이 비교적 크다. 마지막 S-패인은 중복된 영역을 한 번만 집계하여 처리하기 때문에 처리속도가 가장 빠르다.



(그림 4) 튜플의 수에 따른 메모리 사용량

(그림 4)는 튜플의 수에 따른 메모리 사용량을 보여준다. 메모리 사용량은 Normal 일 때 가장 큰 사용량을 보인다. BINT 는 계층구조로 집계정보를 저장해야 하는 비용 때문에 S-패인 보다 메모리비용이 더 들고, S-패인은 중복된 영역을 한 번만 계산하므로 메모리 사용량이 제일 적다.

5. 결론 및 향후 연구

본 논문은 u-GIS 환경의 공간정보를 갖는 연속집계질의간 중복되는 영역의 효율적인 처리를 위한 자원공유기법을 제안하였다. 제안 기법은 공간 데이터 스트림에 적용할 수 있도록 하였다. 우선, 노드 수의 제한이 없고 레벨을 1 로 하는 확장된 R-tree 를 이용하여 공간 집계질의들의 중복되는 영역을 찾는다. 그리고 중복되는 영역을 S-패인 구조로 저장하여 계층별로 모두 집계 값을 유지하는 기존의 기법보다 집계질의 계산의 중복을 줄이고 메모리의 사용량 또한 감소시켰다.

향후 연구로는 R-tree 를 이용하여 중복영역을 줄이는 최소 중복영역 그룹화 알고리즘에 대한 연구가 필요하다.

참고문헌

[1] Babcock, B., Babu, S., Datar, M., Motwani, R. and Widom, J., “Models and Issues in Data Stream Systems.” PODS, 2002

[2] 이충호, 안경환, 이문수, 김주완. u-GIS 공간정보 기술 동향

[3] 백성하, 이동욱, 김경배, 정원일, 배해영. “공간 슬라이딩 윈도우 집계질의의 정확도 향상을 위한 그리드 해쉬 기반의 부하제한 기법”., 한국공간정보시스템학회 논문지 제 11 권 제 1 호, 2009.

[4] A. Arasu, S. Babu, and J. Widom. “The CQL Continuous query Language : Semantic Foundations and Query Execution,” Stanford University Technical Report, 2003.

[5] R. Motwani, J. Widom, A. Arasu, B. Babcock, S. Babu, M. Datar, G. Manku, and C. Olston, J. Rosenstein, and R. Varma, “Query Processing, Resource Management, and Approximation in a Data Stream Management System,” In Proc of CIDR, 2003.

[6] A. Arasu, and J. Widom, “Resource Sharing in Continuous Sliding-Window Aggregates,” In Proc. Of the VLDB, pp.336-347, 2004.

[7] J. Li, and D. Maier, “No Pane, No Gain : Efficient Evaluation of Sliding-Window Aggregates over Data Stream,” SIGMOD Record, pp.39-44, 2005.

[8] Guttman, A., “R-tree: A dynamic index structure for spatial searching,” Proc. Of Intl. Conf. on Management of Data, ACM SIGMOD, 1984.