공간영역질의의 효율적인 연산 공유를 위한 질의영역 밀집도 기반의 그룹화 기법

임정현*, 신숭선*, 백성하*, 이동욱*, 김경배**, 배해영*
*인하대학교 컴퓨터 정보 공학과
**서원대학교 컴퓨터교육과
e-mail : {jhlim, hermit, shbaek, dwlee}@dblab.inha.ac.kr, gbkim@seowon.ac.kr, hybae@inha.ac.kr

Grouping Method Based Query Range Density for Efficient Operation Sharing of Spatial Range Query

Jung-Hyeun Lim*, Soong-Sun Shin*, Sung-Ha Baek, Dong-Wook Lee,
Kyung-Bae Kim, Hae-Young Bae*
*Dept. of Computer Science and Information Engineering, In-ha University
**Dept. of Computer Education, Korea Seowon University

요 약

유비쿼터스 사회를 실현하는 핵심기술인 u-GIS 공간정보 기술은 데이터 스트림 처리 시스템(Data Stream Management System)과 지리정보 시스템(Geography Information System)이 결합된 플랫폼인 u-GIS DSMS 를 요구한다. u-GIS DSMS 는 GeoSeonsor 에서 수집되는 센서 데이터와 GIS 의 공간정보 데이터를 결합하여 처리하는 공간영역질의가 다수 요구된다. 이런 공간영역질의들은 특정 지역에 밀집하게 등록되는 경향이 있으며, 유사한 프리디킷을 가질 가능성이 높다. 이러한 특징은 공간영역질의가 특정 지역에 밀집되면 다수의 비슷한 연산들이 반복적으로 처리하기 때문에 시스템 성능이 저하 될 것이다. 이를 해결하기 위해 영역질의 색인기법 연구가 활발히 진행되고 있다. 그러나 기존의 VCR-Index 와 CQI-Index 기법은 질의영역을 셀 구조나 가상구조로 분할하여 처리하기 때문에 자원 및 연산을 공유 할 수 없어 질의 처리 속도가 현저히 저하되기 때문에 대량의 공간영역질의 처리에는 부적합하다. 그래서 본 논문에서는 공간영역질의의 효율적인 연산 공유를 위한 질의영역 밀집도 기반의 그룹화 기법을 제안한다. 이 기법은 질의영역의 밀집도를 이용하여 공간영역질의들을 그룹화 후 색인을 구성한다. 색인된 영역들의 데이터는 단일 큐로 구성 후 질의들의 프리디킷을 분석하여 자원 및 연산 공유기법을 통해 기존의 기법보다 처리 속도 향상 및 메모리 사용을 감소시켰다.

1. 서론

정보통신기술의 발달로 인하여 유비쿼터스 환경이 도래되었다. 유비쿼터스 환경에서 GIS(Geography Information System)가 결합된 서비스를 제공하기 위해서는 u-GIS 공간정보 기술이 요구된다.

u-GIS 환경에서 다수의 공간영역질의는 GeoSensor 에서 발생하는 센서 데이터와 DBMS 에 저장되어 있는 Historical 데이터를 이용하여 처리한다. 센서 데이터는 시간에 따라 지속적으로 발생되는 데이터 스트림 특성을 지니고 있으며, 자신의 위치 인식 장치를 이용하여 시간에 따라 장소를 이동하는 이동성을 지니고 있다[1]. 또한 Historical 데이터는 공간 및 기타 정보들로 방대한 양을 유지하고 있어 사이즈가 큰 특징을 가진다. 주로 요구되는 서비스(Query)는 특정 지역에 대한 정보를 원하는 경우가 많을 것이다. 예를 들면 “스키장 주변의 팬션을 찾아라” 또는 “콘서트가 열리는 곳에서 주변 도로 차량의 평균속도를 구하라” 등이 있다. 이런 질의들은 유동인구가 많은 특정 지역이나 차량 통행이 혼잡한 지역에서 많이 발생할 것이며, 대부분의 질의들이 유사한 프리디킷을 포함하는 공간영역질의일 것이다.

이런 공간영역질의 처리는 GeoSensor 로 부터 수집되는 정보와 DBMS 에 저장되어 있는 Historical 정보를 결합한 데이터를 해당 질의영역에 접근하여 신속하게 처리해야 하며 대량의 공간영역질의가 연속적으로 발생할 수 있기 때문에 데이터 스트림 기반에서의 공간영역질의의 신속한 처리를 위한 연구가 필요하다. 기존 스트림 기반에서의 연속된 영역질의를 신속하게 처리하기 위해 질의 색인 기법에 대해 활발히 연구를 하고 있다. 다양한 질의색인 기법 중 u-GIS 환경과 비슷한 이동하는 객체에 대해 질의가 고정되어 있는 상황에서의 연속영역질의를 신속하게 처리하기 위한 기법으로 CQI(Cell-based Query Indexing)-Index 와

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

VCR(Virtual Construct Rectangle)-Index 가 있다.

CQI-Index 기법은 그리드 기반의 셀 분할구조를 사용하는 방식으로 질의영역 분할을 편리하고 신속하게 할 수 있으나 셀과 질의영역이 부분적으로 겹치는 부분이 실제 질의영역에 해당되는 데이터인지를 비교하기 위해 정제 단계(Refinement Step)가 필요하기 때문에 연산 비용이 크며, 데이터가 질의영역에 해당되는 셀에 리스트로 연결되어 있는 질의에 의해 처리 되기 때문에 대량의 공간영역질의 처리시 자원 및 연산 공유를 하지 못하여 처리 속도가 크게 저하되는 문제점이 발생한다 [2,3].

VCR-Index 기법은 가상분할 구조를 사용하는 방식으로 CQI-Index 와는 달리 질의영역을 미리 정의된 VCR(Virtual Construct Rectangle) 을 이용하여 분할하는 색인 방식이다. 이 기법은 CQI-Index 기법과 달리 부분적으로 겹쳐지는 영역들을 없애 신속한 검색을 보장했지만 모든 질의영역을 VCR 로 분할하기 때문에 공간영역질의가 많아지면 영역을 분할한 VCR 이 많아져 관리 비용이 커지게 되고 분할된 영역 마다 질의 ID 가 저장되어 있기 때문에 대량의 공간영역질의 처리시 자원 및 연산 공유를 하지 못하여 메모리 낭비의 문제점과 속도 저하의 문제점이 발생한다[4,5].

기존의 문제점을 해결하기 위해서 본 논문은 공간영역질의의 효율적인 연산 공유를 위한 질의영역 밀집도 기반의 그룹화 기법을 제안한다. 기존의 기법들은 데이터가 수집되면 미리 색인된 영역을 검색하여 해당영역에 저장된 질의 처리를 하기 때문에 자원 및 연산 공유 기법을 질의 처리에 적용 할 수 없다. 자원 및 연산공유 기법을 질의 처리에 적용하기 위해서는 질의영역의 밀집도를 기반으로 하여 질의영역을 그룹화 한다. 그룹화 된 영역에 해당하는 데이터가 수집되면 그룹마다 단일 버퍼에 담아 해당되는 질의들의 연산을 분석하여 자원 및 연산 공유를 활용함으로써 신속하게 질의 처리가 되며, 전체 메모리 사용이 감소된다.

본 논문의 구성은 다음과 같다. 2 절에서는 연속 질의 처리를 위한 연산 공유 및 스케쥴링과 영역질의 색인 기법에 대해 설명한다. 3 절에서는 본 논문에서 제안하는 밀집된 질의영역에 대한 그룹화 기법 및 자원 및 연산 공유 기법을 서술한다., 4 절에서는 제안한 기법의 성능측정 결과를 평가하고, 5 절에서는 본 논문의 결론 및 향후 연구를 보인다.

2. 관련연구

2.1 DSMS 에서 연속 질의 처리를 위한 연산 공유 및 스케쥴링 기법

데이터 스트림 환경에서 연속 질의 처리를 위한 연구가 활발히 진행 중에 있다. 다양한 기법들 중에서 연산 공유 및 스케쥴링을 이용한 시스템들로 대표적인 데이터 스트림 처리 시스템으로 STR EAM(The Stanford Steam Data Manager) 과 Aurora 등이 있다.

STREAM 시스템은 연속 질의 처리를 위해 질의의 프리디킷을 연산자 단위로 분리 후, 연산비용이 최소화 될 수 있도록 연산자 체인을 구성하여 질의 계획을 세운다. 질의 계획은 연산자 공유, 연산 처리결과를 공유 하는 등 다양한 방법을 통해 메모리 관리를 효율적으로 하며, 신속한 질의 처리를 할 수 있도록 한다. 본 논문에서 제안하는 기법은 STREAM 시스템의 자원공유 및 연산자 공유를 활용 한다[6].

Aurora 시스템은 연속 질의 처리를 위해 질의와 연산자 단위로 처리 연산을 분할하여 질의를 동시에 처리하고 중요한 데이터를 정적 저장 장소에 버퍼링하는 시스템 구조를 이용하여 신속한 질의처리를 제공한다. 또한 최적화를 위해서 연산자 스케쥴링 및 연산자 결합 및 재정렬을 통해서 처리 효율을 높인다. 본 논문에서 제안하는 기법은 Aurora 시스템의 연산자 결합 및 재정렬을 활용한다[7].

2.2 이동객체에 대한 영역질의 색인 기법

u-GIS 와 유사한 조건의 질의 색인 기법에 이동객체에 대한 연속 영역질의를 메모리 상에서 처리하는 기법으로 분할구조와 셀 분할구조를 사용하는 다차원 질의 색인 구조로 VCR(Virtual Construct Rectangle) Index 와 CQI(Cell-based Query Indexing) Index 가 있다.

셀 분할구조로 대표적인 CQI-Index 기법은 그리드 기반의 셀 분할구조를 사용하는 방식으로 전체 영역을 일정 크기의 셀로 분할하고 각 셀은 연관되는 노드를 가진다. 셀은 일정한 크기여서 질의 영역을 분할 삽입 시 셀에 완전히 겹쳐지는 부분과 부분적으로 겹치는 부분이 존재하게 된다. 따라서 CQI-Index 는 노드에 두 개의 리스트를 유지한다. 하나는 Full List 로 분할된 영역이 셀에 완전히 겹쳐진 경우를 위한 것이며, 나머지 하나는 Part List 는 분할된 영역이 셀에 부분적으로 겹쳐진 경우를 위한 리스트이다. 검색 시 Part List 를 탐색하여 얻은 결과인 질의 셋이 실제로 이동체를 포함하는지를 비교하기 위한 정제 단계가 필요한 특징을 가지게 되는데 영역질의가 많아지게 되면 Part List 가 많아져 정제 단계 처리 비용이 높아 처리 성능을 저하 시키는 문제점이 발생한다.

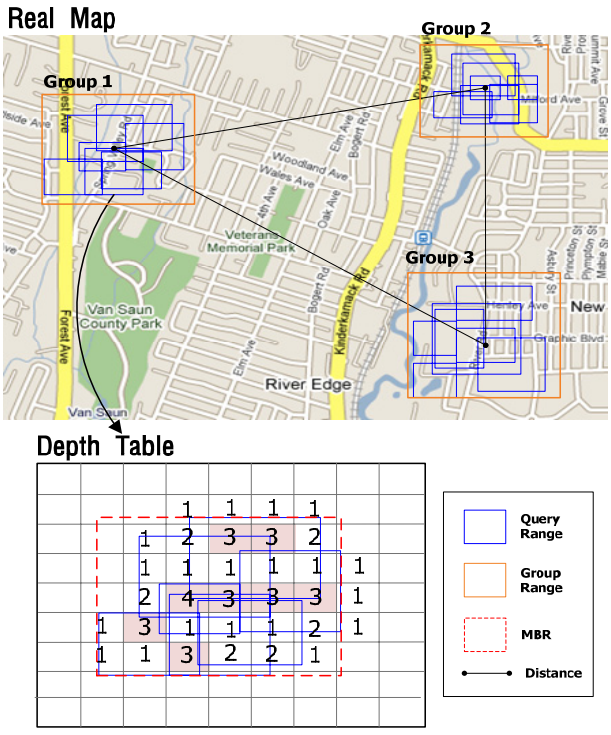
가상분할 구조로 대표적인 VCR-Index 기법은 영역질의의 프리디킷이 2 차원 사각형으로 표현됨을 활용, 미리 정의된 VCR(Virtual Construct Rectangle)을 이용하여 분할하여 해당 구조와 연관되는 노드의 ID 리스트에 질의 ID 를 삽입하는 방식이다. 질의영역을 분할할 때 CQI-Index 와는 달리 부분적으로 겹쳐지는 부분이 없기 때문에 빠른 검색이 가능하다. 하지만 각각의 질의영역을 VCR 로 분할하기 때문에 연속적으로 많은 공간영역질의 처리가 요구되면 분할된 영역이 많아져 관리 비용이 커지게 되고 겹쳐지는 영역에 대해서 반복적으로 연산을 수행하기 때문에 공간 낭비 및 처리 성능을 저하 시키는 문제점이 발생한다.

3. 본론

3.1 Depth Table 을 이용한 밀집된 질의 영역 그룹화 기법

대량의 공간영역질의를 신속하게 처리하기 위해서는 자원 및 연산 공유 기법을 적용해 질의 처리를 해야 한다. 하지만 자원 및 연산 공유 기법을 적용하려면 질의에 해당되는 데이터들이 단일 큐로 구성되어 있어야 가능하다. 전체 공간영역을 단일 큐로 구성하기에는 메모리 공간이 부족하게 되어 사실상 불가능하며, 기존의 기법에서도 질의 영역을 분할하여 질의 처리를 했기 때문에 여러 개의 큐에 존재하는 데이터들을 공유하거나 질의들을 그룹화 하여 처리 할 수가 없어 메모리가 낭비 되고 처리 속도도 느리게 되는 문제점이 발생한다. 이런 문제점을 해결하고 자원 및 연산 공유 기법을 적용하기 위해 밀집된 질의영역을 그룹화 하여 그룹 된 영역만큼의 데이터만 큐에 삽입하여 사용한다면 메모리 사용의 문제점과 자원 및 연산 공유의 문제점을 해결 할 수 있다. 본 절에서는 Depth Table 을 이용한 밀집된 질의 영역 그룹화 기법을 설명한다.

질의가 밀집된 지역을 그룹화 하기 위해서는 전체 공간영역을 고정 그리드 기반의 셀로 분할 후 질의영역에 해당하는 셀의 카운터를 증가시킨다. 카운터가 가장 큰 셀들부터 체크를 하여 가장 큰 카운터의 90%의 카운터까지의 셀들을 체크를 한다. 체크된 셀들의 거리를 측정하여 거리가 가까운 것끼리는 같이 그룹화를 하고 거리가 먼 것은 따로 그룹화를 실시한다. 그룹화는 체크된 셀에 포함되는 질의영역을 검색한다. 검색된 영역을 포함할 수 있는 MBR(Minium Bound Rectangle)로 그룹을 구성하고 포함 되는 질의를 리스트에 저장한다.



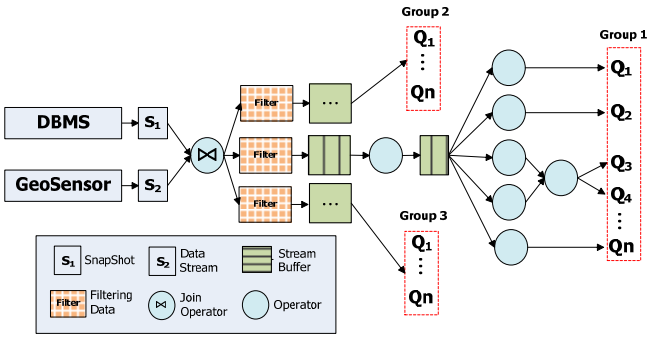
(그림 1) Depth Table 을 이용한 질의영역 그룹화

예를 들면, (그림 1)는 실제 지도상에서 공간영역질의가 Real Map 에서와 같이 요구되었을 때 Depth Table 를 이용한 그룹화 과정을 나타낸다. Real Map 전체 영역을 셀 기반으로 분할 하여 Depth Table 을 이용하여 각 셀들간의 거리를 측정한 후 영역별로 나눈다. 나뉘어진 영역마다 그룹 영역을 구성한다. (그림 1) 에서 Depth Table 의 가장 큰 셀의 카운터는 4 이다. 4 의 90%까지 셀을 체크해야 하므로 카운터의 크기가 4, 3 인 셀에 포함되는 질의영역을 모두 포함하는 MBR(Minimum Bound Rectangle) 을 구성한다. MBR 로 그룹화된 영역은 점선과 같이 나타낼 수 있다.

3.2 그룹화된 영역질의의 신속한 처리를 위한 자원 및 연산 공유 기법

그룹화된 영역질의의 신속한 처리를 위한 자원 및 연산 공유 기법은 1 차적으로 필터링 된 데이터들 중 질의 처리에 불필요한 데이터를 제거 해야 한다. 또한 여러 질의들이 데이터를 공유해서 사용할 수 있기 때문에 질의들의 프리디킷의 연산들을 분리하여 처리 비용 및 연산 공유 여부 등에 관한 연산 스케쥴링을 구성해야 한다. 연산 스케쥴링을 구성하게 되면 질의 처리 성능을 향상 시킬 수 있기 때문에 그룹화된 영역질의 신속한 처리를 위한 자원 및 연산 공유 기법에 대해 설명한다.

밀집된 지역에서 실시간으로 요구된 공간영역질의들의 프리디킷은 비슷한 패턴을 가지거나 동일한 경우가 많다. 이 점을 활용하여 질의의 프리디킷을 연산 별로 분리 후 처리 비용과 필요한 데이터가 무엇인지를 검사한다. 필요한 데이터가 같은 질의끼리 그룹을 지어 동시에 질의 처리를 하도록 한다. 연산자 스케쥴링은 불필요한 데이터를 최대한 제거하기 위해 비공간 연산을 선 처리하고 연산 비용이 큰 공간연산을 불필요한 데이터가 제거된 데이터를 가지고 연산함으로써 처리 성능을 향상 시킬 수 있다. 이런 질의 처리의 흐름도를 나타내면 (그림 2)와 같이 나타낼 수 있다.



(그림 2) 공간영역질의 처리 흐름도

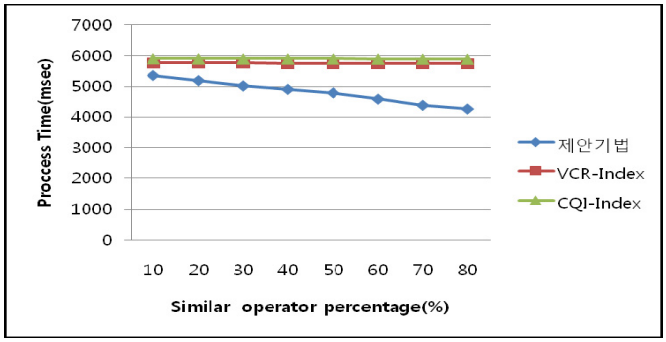
4. 성능평가

4.1 실험환경

성능평가에 사용된 시스템은 Intel® Pentium® 4 3.0GHz CPU Chipset 을 사용하였으며, 메모리는 2GB, 운영체제는 Windows XP 상에서 구동하였다. 평가 방법은 타 기법과 본 논문이 제안하는 집중된 공간영역질의의 그룹화를 통한 연산자 스케쥴링이 특정 지역에 집중된 공간영역질의에 대한 처리 성능을 비교하여 본 기법의 우수성을 증명한다. 비교를 위해 두 기법의 시뮬레이터를 제작하여 평가 하였으며, 시뮬레이터는 Visual C++ 2005 기반으로 작성하였다. 평가에 사용된 데이터 셋은 분할된 연산자로 구성되었으며 각 연산자의 평균 처리 값은 임의로 지정하여 사용하였다.

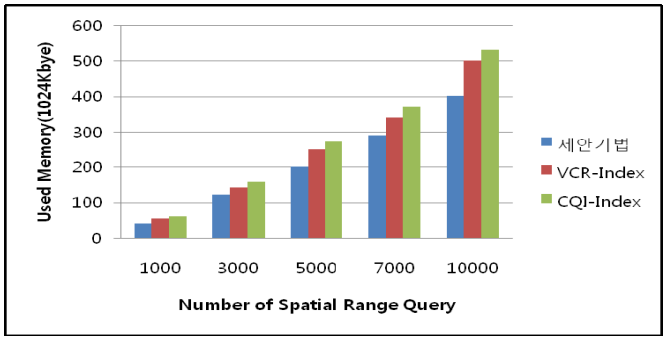
4.2 성능평가

본 성능평가 에서는 본 논문이 제안한 기법과 기존의 VCR-Index 와 CQI-Index 의 유사 연산 비율에 따른 처리 시간에 대해서 측정하였다.



(그림 3) 유사 연산 비율에 따른 처리 시간 비교

(그림 3)는 공간영역질의 5000 개 처리 시 유사연산자 비율에 따른 처리 시간 그래프이다. CQI-Index 와 VCR-Index 는 비율에 관계 없이 일정한 처리 시간을 보였다. 집중된 공간영역질의의 그룹화를 통한 연산자 스케쥴링은 연산자 비율이 높아질 수록 처리 시간이 현저하게 줄어든 것을 알 수 있다. 또 다른 성능 평가로 공간영역질의 수에 따른 메모리 사용량을 측정하였다.



(그림 4) 공간영역질의 처리 시 메모리 사용 비교

(그림 4)는 공간영역질의 개수에 따른 메모리 사용량 그래프이다. 기존 기법에 비해 메모리 사용량이 줄어든 것을 알 수 있다.

5. 결론 및 향후연구

본 논문에서는 u-GIS 환경에서 연속적이며, 대량의 공간영역질의의 신속한 처리를 위한 공간영역질의의 효율적인 연산 공유를 위한 질의영역 밀집도 기반의 그룹화 기법을 제안하였다.

본 기법은은 특정지역에 연속적으로 다수의 공간영역질의가 요구되었을 시에 밀집된 지역을 중심으로 질의영역을 MBR 로 그룹화 하여 데이터 스트림을 필터링한다. 필터링 된 데이터들을 그룹에 속한 질의들의 연산을 분리하여 자원 및 연산 공유 기법을 적용해 질의 처리 성능을 향상시켰다.

향후 연구로는 질의 영역에 포함되지 않는 영역을 최소화 하는 그룹화 기법에 대한 연구 및 성능향상을 위한 연산자 스케쥴링에 대한 연구를 진행할 것이다.

참고문헌

[1] 이충호, 안경환, 이문수, 김주완, “u-GIS 공간정보 기술 동향”, 전자통신동향분석, ETRI

[2] D. V. Kalashnikov, S. Prabhakar, W. G. Aref, S. E. Hambrusch, “Efficient Evaluation of Continuous Range Queries on Moving Objects”, In Proc. Of 13th Intl. Conference on Database and Expert System Applications - DEXA, 2002

[3] Kun-Lung Wu, Shyh-Kwei Chen, and Philip S. Yu, “Processing Continual Range Queries over Moving Objects Using VCR-Based Query Indexes”, Proc. IEEE International Conference on Mobile and Ubiquitous Systems : Networking and Services, pp.226-235, 2004

[4] D. V. Kalashnikov, S. Prabhakar, and S. E. Hambrusch, “Main Memory Evaluation of Monitoring Queries Over Moving Objects”, Distributed and Parallel Databases, Vol.15, No.2, pp.117-135, 2004

[5] Kun-Lung Wu, Shyh-Kwei Chen and Philip S. Yu, “VCR Indexing for Fast Event Matching for Highly-Overlapping Range Predicates”, SAC`04 March 14-17 ACM, 2004

[6] A. arasu, B. Babcock, S. Babu, M. Datar, K. Ito, R. Motwani, I. Nishizawa, U.Srivastava, D. Thomas, R.Varma and J.Widom “STREAM : The Stanford Stream Data Manager.”, IEEE, 2003

[7] Daniel J. Abadi, Don Carney, Ugur Cetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul, Stan Zdonik, “Aurora : a new model and architecture for data stream management”, The VLDB journal, 2003