데이터 스트림에서 공간질의의 영역 겹침을 이용한 우선순위 기반의 부하 분산 기법

김호*, 백성하*, 이연*, 이동욱*, 정원일**, 배해영*
*인하대학교 정보공학과
**호서대학교 정보보호학과
e-mail : {hokim, shbaek, leeyeon, dwlee}@dblab.inha.ac.kr, wnchung@hoseo.edu, hybae@inha.ac.kr

Priority based Load Shedding Method using Range Overlap of Spatial Queries on Data Stream

Ho Kim*, Sung-Ha Baek*, Yan Li*, Dong-Wook Lee*,
Weon-Il Chung**, Hae-Young Bae*
* Dept. of Computer Science and Information Engineering, Inha University
** Dept. of Information Security Engineering, Hoseo University

요 약

u-GIS 환경에서 발생하는 시공간 데이터는 지속적으로 발생하는 데이터 스트림의 특성을 갖으며, 그런 특성으로 인하여 데이터 발생량이 급격히 증가함에 따라 데이터 손실 및 시스템 성능 저하현상이 발생한다. 이를 해결하기 위해 부하 분산 연구들이 활발히 진행되어 오고 있다. 그러나 기존의 연구 방식인 랜덤 부하 분산 방식과 의미적 부하 분산 방식은 현 u-GIS 환경에서 부하 분산 속도 및 질의 결과의 정확도 측면에 만족스럽지 못한 결과를 준다. 그래서 본 논문에서는 우선순위를 이용한 차등적 부하 분산(DLSM : Different Load Shedding using MAP table)기법을 제안한다. DLSM 기법은 등록된 공간질의의 공간연산을 통해 영역의 우선순위를 미리 부여하고, 데이터가 발생하여 질의 처리기로 유입되기 전 우선순위를 파악한다. 데이터는 우선순위 단계에 따라 유입량을 확인 후 삭제 여부가 결정된다. 결과적으로 부하 분산 속도와 질의 결과의 정확도를 향상시켰다.

1. 서론

유비쿼터스 환경의 핵심 기반 기술로 대두되고 있는 u-GIS 공간정보 기술은 시간에 따라 공간적인 위치가 포함된 동적인 데이터를 GeoSenSor(e.g. RFID 리더, 모바일 RFID 리더, 센서노드, CCTV 등)로부터 수집한다. 다양한 형태로 수집되는 이 데이터들은 시간에 따라 지속적으로 발생되는 데이터 스트림의 특성을 갖는다[8]. 데이터 스트림은 데이터의 지속적인 발생과 빠른 입력 속도의 특성때문에 유한하게 한정되어 있는 메모리 공간과 입력 데이터를 실시간으로 처리하기 부족한 환경에서는 일부 데이터가 처리되지 못하고 소실되거나 시스템 성능 저하현상이 나타나는 문제점을 가진다. 이런 문제점을 해결하기 위해 크게 입력속도를 기반으로 한 랜덤 부하 분산 방식과 데이터의 중요도를 기반으로 한 의미적 부하 분산 방식이 연구되고 있다[2,9].

랜덤 부하 분산 방식은 삭제될 데이터를 랜덤 연산에 의해 빠르게 선택하여 부하 분산을 하는 장점을 갖지만, 데이터의 중요도를 고려하지 않기 때문에 중요 데이터가 삭제되어 질의 결과의 정확도를 감소시킬 수 있다. 의미적 부하 분산 방식은 데이터 및 질의에 우선순위를 부여하여 데이터의 중요도를 판단한다. 중요도가 낮은 데이터는 우선적으로 삭제하고 중요도가 높은 데이터를 이용하여 질의 결과의 정확도를 향상시킨다. 그러나 데이터의 중요도를 반영하기 위한 연산과정이 복잡하고 각 데이터마다 반복적으로 수행되어 부하 분산을 위한 수행 속도가 느리다는 단점을 갖는다. 이에 반해 u-GIS 환경에서의 공간 연속 질의는 위치 정보가 포함된 데이터를 이용하여 질의 처리되어야 하며, GeoSenSor 로부터 빠르게 수집되는 많은 양의 데이터들을 처리할 수 있는 부하 분산 방식이 요구된다. 즉, 랜덤 부하 분산 방식의 빠른 부하 분산 속도와 의미적 부하 분산 방식의 높은 질의 결과의 정확도를 동시적으로 요구한다.

본 논문에서는 부하 분산 속도와 질의 결과의 정확도를 향상시키기 위해, 공간질의의 영역 겹침에 의한 데이터 우선순위를 부여하고 부여된 우선순위를 고려한 차등적 삭제 기법을 제안한다. 공간연속질의에 제시된 공간연산을 분석하여 영역의 겹침 정도에 따라 공간의 중요도를 부여하고, u-GIS 환경에서 발생하는 시공간 데이터들의 위치 정보를 이용하여 우선순위 단계를 확인한다. 우선순위 단계는 질의를 분석하여 미리 부여한 것으로 시공간 데이터의 위치 정

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

보와 비교 연산만을 통해 데이터의 우선순위 정도를 빠르게 확인 가능하다. 이 후, 파악된 우선순위 정도에 따라 차등적으로 데이터를 삭제하여 부하를 분산한다. 결과적으로 랜덤 부하 분산 방식의 문제점으로 지목했던 정확도는 시공간 데이터의 우선순위를 부여하여 중요 데이터를 판별함으로써 향상시킬 수 있었으며, 의미적 부하 분산 방식의 문제점으로 지목했던 복잡한 알고리즘에 의한 부하 분산 속도 저하 현상은 최소한의 연산과정을 통해 부하 분산 속도 향상시킬 수 있었다.

2. 관련연구

2.1 랜덤 부하 분산 기법

랜덤 부하 분산 기법은 유입되는 데이터와 질의에 대해 우선순위를 부여하지 않고 무작위로 데이터를 선택하여 삭제한다[4]. 랜덤 부하 분산 알고리즘은 데이터의 중요도를 고려하지 않고 랜덤 연산으로만 선택하기 때문에 빠른 부하 분산 속도가 장점이다. 그러나 중요도가 높은 데이터들이 삭제될 경우 질의 결과에 대한 정확도가 현저히 감소하여 신뢰성을 확보하지 못하는 문제점을 갖는다.

특히, 본 논문에서 거론한 u-GIS 환경에서 발생하는 시공간 데이터의 위치 정보를 고려하지 않고 랜덤하게 선택하여 삭제하기 때문에 중요 위치 정보를 포함한 데이터가 알고리즘에 의하여 삭제되거나 손상되어 질의 결과의 정확도를 확보하지는 못하는 상황이 발생한다.

2.2 의미적 부하 분산 기법

데이터 스트림에서 연속 질의를 위한 우선순위 기반의 의미적 부하 분산 기법을 제안하였다[9]. 등록되는 연속질의에 우선순위를 할당하면 부하 제한기에 의해 질의 우선순위 정도를 파악한다. 낮은 우선순위의 질의가 액세스하는 데이터를 우선적으로 삭제하여, 높은 우선순위의 질의 처리에 대해 부하가 발생하지 않도록 사전에 제한하는 기법이다. 이 기법은 우선순위가 높은 질의들의 정확도를 향상시키고 QoS 를 지원하는 장점을 갖는다. 그러나 지속적으로 높은 우선순위의 질의가 등록되면 낮은 우선순위의 질의는 기아현상이 발생한다. 연속질의가 등록되면 데이터영역 설정 및 데이터 영역 우선순위 배정에 대한 알고리즘이 반복적으로 실행되기 부한 분산 속도가 저하되는 현상을 보인다. 또한 질의의 우선순위 정도가 데이터 영역에도 상속되기 때문에 사용하지 않는 데이터 영역이 발생한다. 하나의 데이터 스트림의 동일한 우선순위를 갖는 연속질의가 등록되었을 때, 데이터 역시 동일한 우선순위의 데이터가 유입되어 낮은 우선순위의 데이터가 존재하지 않아 우선적으로 삭제할 수 있는 데이터가 존재하지 않게 된다.

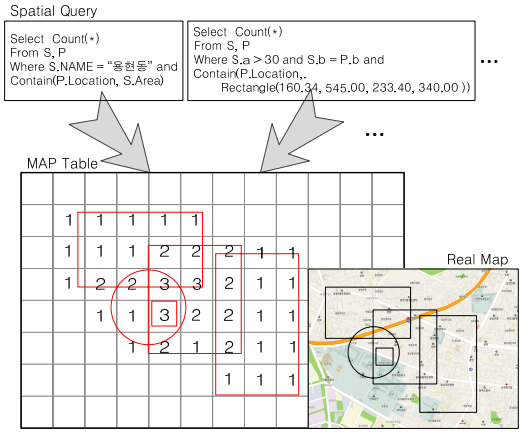
본 논문의 u-GIS 환경에서는 하나의 공간연속질의만으로도 다양한 GeoSenSor 로부터 많은 양의 시공간 데이터가 빠르게 유입되어 부하가 발생할 경우 부하 분산이 어렵다. 더 큰 문제점은 시공간 데이터는 문자, 숫자, 이미지 등의 다양한 형식으로 이루어져 데이터 영역을 설정할 수 없는 데이터 유입되면 우선순위 부여 자체가 불가능해져 데이터 손실 및 시스템 저하현상을 초래한다.

3. 본론

u-GIS 환경에서 위치·시간 정보를 포함한 시공간 데이터는 GeoSenSor 로부터 수집되어 처리된다. GeoSensor 로부터 수집되는 시공간 데이터는 위치 정보 이외 문자, 숫자, 이미지 등의 다양한 형태를 갖는다. 이런 다양한 형태의 시공간 데이터를 빠르고 정확하게 처리하기 위하여 MAP 테이블 생성 알고리즘과 MAP 테이블을 이용한 차등적 부하 분산 알고리즘을 결합시킨 DLSM 기법을 설명한다. DLSM 기법은 데이터의 공통 요소인 위치 정보를 활용하여 공간연속질의에 우선순위를 부여하고 부여된 우선순위를 기반으로 차등적으로 삭제 여부를 결정지어 부하 분산을 수행한다. 단, 하나의 데이터 스트림은 한정적인 공간을 대상으로 균일한 그리드 셀로 구성하여 공간연속질의에 우선순위를 부여한다.

3.1 우선순위 MAP 테이블 생성

u-GIS 환경에서 발생하는 시공간 데이터의 부하를 분산하기 위하여 공간연속질의의 영역 겹침에 의한 우선순위 MAP 테이블 생성 알고리즘을 설명한다. 본 알고리즘은 의미적 부하 분산 방식의 복잡한 연산과정으로 인한 부하 분산 속도 감소를 문제를 해결할 수 있다.



<그림 1> MAP 테이블 우선순위 부여

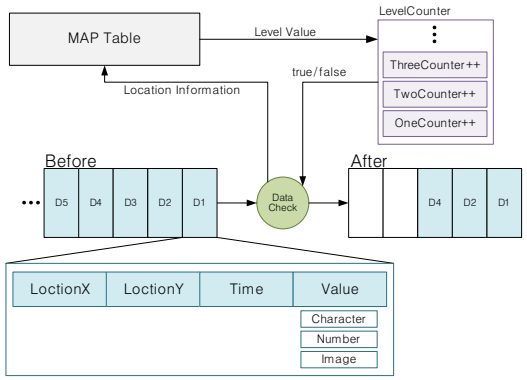
<그림 1>과 같이 등록된 공간연속질의에서 제시한 공간 영역을 균일한 그리드 셀로 구성 후, 해당 셀의 우선순위 카운터를 증가시켜 영역의 중요도를 반영한다. 공간연속질의는 기본적으로 SQL 문 형태를 갖추고 있으며 영역범위를 설정하는 공간연산(CONTAIN, OVERLAP, EQUEALS, DISJOINT 등)을 포함한다. 공간연산은 MAP 테이블에 질의 영역을 설정하기 위하여 등록된 질의의 구문을 분석하고 질의에 포함된 공간연산의 범위를 추출한다. MAP 테이블에서는 추출된 공

간 영역을 토대로 일치하는 셀을 검색하고 검색된 셀에는 우선순위 카운터를 증가시킨다. 우선순위 카운터의 기본 값은 0 이며, 공간질의의 공간 연산의 영역 설정수만큼 우선순위를 N 단계까지 증가한다. 본 논문에서는 알고리즘에 설명을 돕기 위하여 최대 10 단계의 우선순위로 가정하여 중요도를 반영한다.

등록된 공간연산질의의 삭제 요청이 들어오면, 삭제 요청된 질의의 영역범위를 분석하고, 분석한 영역의 범위를 MAP 테이블에서 찾는다. 해당 셀의 우선순위 카운터를 감소시켜 공간질의의 해당 영역의 중요도를 낮추고, 정상적으로 질의를 삭제한다.

3.2 MAP 테이블 이용한 차등적 부하 분산(DLSM)

본 절에서는 MAP 테이블을 이용하여 데이터의 우선순위를 부여 받고, 그 우선순위에 따라 차등적으로 부하 분산하는 알고리즘을 설명한다. 본 알고리즘은 랜덤 부하 분산 기법에서 발생하는 질의 결과의 정확도를 향상시킬 수 있다. 또한, 데이터의 다양한 형태(문자, 숫자, 이미지 등)로 인하여 데이터 영역을 나눌 수 없는 문제점을 위치 정보를 이용하여 우선순위를 부여하고 우선순위에 의한 부하 분산을 수행하여 해결한다.



<그림 2> DLSM 기법의 구성도

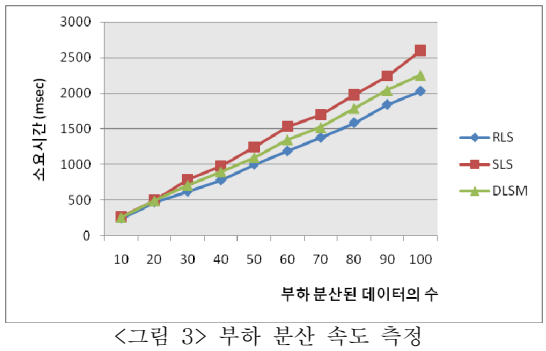
<그림 2>와 같이 부하 분산은 질의가 처리되기 전인 데이터 스트림에서 수행한다. 데이터가 질의 처리를 위해 데이터 스트림에서 입력되어지는 시점은 연산자들을 공유하지 않는 질의를 위한 최적의 시점이다[2]. 이 때 시공간 데이터의 위치 정보를 분석한다. 분석된 데이터는 MAP 테이블로부터 우선순위 단계를 부여 받고 우선순위 정도에 따라 LevelCounter 를 통해 데이터의 유입량을 체크하여 삭제 여부를 결정한다. 여기서 LevelCounter 는 질의 처리기로 삽입되어진 데이터의 수를 체크하는 변수로, 기본 값은 0 이며, 데이터의 우선순위 단계의 LevelCounter 를 증가시키고 해당 우선순위 단계와 카운터의 값이 같은지를 비교한다. 증가한 카운터의 값이 우선순위 단계보다 작다면 데이터는 정상적으로 유입되지만, 그 값이 같다면 데이터는 삭제되고 카운터의 값은 0 으로 재설정된다.

본 논문에서는 알고리즘의 설명을 돕기 위하여 우선순위의 정도를 10 단계로 가정하였다. 부여되는 우선순위는 데이터의 중요도를 나타내는 동시에 데이터 스트림의 삽입할 수 있는 개수를 의미한다. 우선순위가 3 인 데이터의 경우 3 의 우선순위를 갖는 데이터는 3 개까지 삽입되며 4 번째 삽입되는 데이터는 삭제한다. 같은 방법으로 10 의 우선순위의 데이터라면, 10 개의 데이터의 삽입 후 11 번째 데이터는 삭제한다. 예를 들어 데이터(Di)의 우선순위가 다음과 같다면, {D1(3), D2(2), D3(1), D4(5), D5(5), D6(2), D7(1), D8(2), D9(3), …} 1 의 우선순위 데이터 중 두 번째로 유입되는 D7 과 2 의 우선순위 데이터 중 세 번째로 유입되는 D8 이 부하 분산의 대상이 된다. 이 같은 방법은 높은 우선순위의 데이터를 질의처리에 보다 많은 수를 반영하여 정확도 향상에 목적을 두었다. 더불어 공간질의는 특정 중요 지점(Point)이 아닌 영역(Area)을 제시하는 질의이므로 낮은 우선순위의 데이터일지라도 일부는 질의처리에 반영하도록 한다.

4. 성능평가

본 논문에서 제안한 DLSM 기법에 대한 성능을 평가하기 위하여 기존 기법인 랜덤 부하 분산 기법(RLS)과 의미적 부하 분산 기법(SLS)에 대한 테스트를 수행하였다. 테스트에 사용된 시스템의 환경은 Intel Pentium 4 CPU 3.0GHz, 4GB Memory, 160GB Hard 이다. 테스트를 위하여 MAP 테이블의 셀에 최대 10 단계의 우선순위를 임의적으로 부여하되, 주변 셀과 비슷한 우선순위를 가질 수 있도록 하여 질의 영역에 의해 묶인 형태를 갖도록 하였다. 데이터는 데이터 생성기에 의해 시간과 위치 정보가 포함되어 지속적으로 생성한다. 생성되는 시공간 데이터의 스키마는 <long LocationX, long LoactationY, int DateInfo, int TimeInfo, int RandomValue>와 같으며 크기는 28byte 로 고정시켰다.

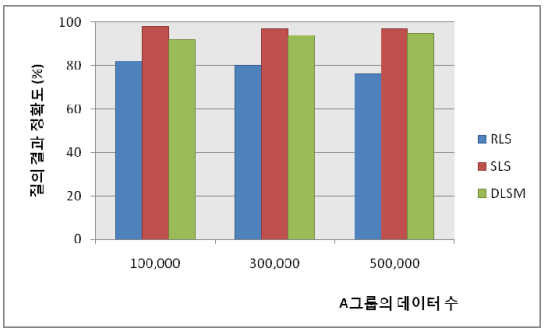
먼저 부하 분산 속도를 측정하기 위해 다음과 같이 진행하였다. 지속적으로 데이터를 삽입하면서 임의의 데이터 수만큼 부하 분산을 수행하는데 소요된 시간을 측정한 것으로 3 가지 기법의 시간차를 측정한다. 단, 데이터 스트림에 일정량의 데이터를 삽입 후 수행하였다.



<그림 3> 부하 분산 속도 측정

<그림 3>와 같이 DLSM 기법은 의미적 부하 분산 기법(SLS)의 부하 분산 속도를 향상시켜 랜덤 부하 분산 속도(RLS)의 근접하였다. DLSM 기법은 위치 정보만으로 우선순위를 파악하고 데이터 스트림으로 유입되어진 데이터의 개수는 별개의 카운터 변수의 의해 관리되어지기 때문에 짧은 연산과정에 의하여 부하 분산의 수행 속도가 소폭 향상되었다.

다음은 질의 결과의 정확도를 측정하기 위하여 다음과 같이 진행하였다. 1,000,000 개의 데이터를 미리 생성하고 데이터의 영역을 두 그룹으로 나누어 삽입을 수행한다. A 그룹의 데이터의 RandomValue 에는 30 보다 큰 값을, 나머지 B 그룹 데이터에는 30 을 포함한 작은 값을 임의로 할당한다. 질의는 "남구 내에서 온도가 30 도가 넘는 건물의 수를 구하라." 설정하였으며, 질의가 설정한 강남구 영역 안에서만 RandomValue 가 30 이 넘는 A 그룹이 생성되도록 하였다. B 그룹의 경우는 질의 영역과 영역 밖에서 모두 발생한다. 즉, A 그룹으로 설정한 데이터의 개수를 통해 부하 분산으로 발생하게 될 손실율 측정하여 정확도를 나타내었다.



<그림 4> 질의 결과의 정확도 측정

<그림 4>와 같이 랜덤 부하 분산 기법(RLS)의 질의 결과의 정확도는 매 실험 때마다 큰 차이를 보였다. 이에 반해 DLSM 기법은 데이터의 우선순위 단계에 의해 데이터 삭제 여부 판단되기 때문에 높은 우선순위의 데이터가 질의 처리에 보다 많이 반영되고 낮은 우선순위의 데이터는 보다 적게 반영되어 질의 결과의 정확도가 향상되었다.

5. 결론 및 향후 연구

지속적으로 발생하는 시공간 데이터의 부하를 분산하기 위하여 공간 연산의 영역 겹침에 의한 MAP 테이블 생성 알고리즘과 우선순위 MAP 테이블을 이용한 차등적 삭제 알고리즘을 결합한 DLSM 기법을 제안하였다. DLSM 기법은 등록된 공간질의의 공간연산을 분석하여 우선순위 MAP 테이블을 생성하며, 데이터들은 MAP 테이블을 이용하여 우선순위를 파악하고 부하 분산을 위해 삭제 여부를 결정짓기 때문에 최소한의 연산과정으로 부하 분산 속도를 향상시켰다. 또한 데이터의 중요도를 충분히 반영하여 높은 우선순위의 데이터는 보다 많이, 낮은 우선순위의 데이터는 보다 적게 질의 처리되도록 하여 질의 결과의 정확도를 향상시켰다. 성능평가를 통하여 u-GIS 환경에서 기존의 연구 기법의 단점을 최소화하고, 부하 분산 속도와 질의 결과의 정확도를 향상하였음을 보였다.

향후 연구로는 등록된 질의의 개수가 많은 경우와 추가적으로 질의가 등록되는 경우에 우선순위의 단계를 그룹화하여 한정된 우선순위의 단계를 설정되도록 하는 것으로 그룹화할 우선순위의 범위 설정 및 몇 단계의 그룹 우선순위를 지정할 것인지에 대한 최적화 연구를 진행할 것이다.

참고문헌

[1] B.Babcock, S. Babu, M. Datar, R. Motwani, and j. Widow, "Models and Issues in Data Stream Systems," Invited paper in Proc of PODS, 2002.

[2] N. Tatbul, U. Cetintemel, S. Zdonik, M. Chemiack, and M. Stonebraker, "Load Shedding in a Data Stream Manager," In Proc. Of the 29th VLDB Conf. pp.309-320, 2003.

[3] N. Tatbul, and S. Zdonik, "Window-Aware Load Shedding for Aggregation Queries over Data Streams," VLDB Sep, 2006.

[4] B. Babcock, M. Datar, and R. Motwani, "Load shedding for Aggregation Queries over Data Streams," Proc. Of the 20th ICDE, pp.1-12, 2004.

[5] D. Abadi, D. Carney, U. Cetintemel, M. Cherniack, C. Convey, S. Lee, M. Stonebraker, N. Tatbul, and S. Zdonik, "Aurora : A New Model and Architecture for Data Stream Management," VLDB J., vol.12, no.2, pp.120-139, 2003.

[6] N.Tatbul and S. Zdonik, "Dealing with Overload in Distributed Stream Processing Systems," In IEEE International Workshop on Networking Meets Databases*(NetDB'06)*, 2006.

[7] R. Avnur and J. M. Hellerstein, "Eddies : Continuously Adaptive Query Processing," In Proc. Of the ACM SIGMOD *International Conference on Management of Data*, pp.261-272, 2000.

[8] 이충호, 안경환, 이문수, 김주완, "u-GIS 공간정보 기술 동향," 전자통신동향분석, ETRI, 2007.

[9] 박재석, 조행래, "데이터스트림에 대한 연속 질의를 위한 우선순위 기반의 의미적 부하 제한," 데이터베이스연구 제 21 권 제 1 호, 2005.