병렬 공간데이터베이스 시스템에서 공간 정보 처리 방안

김진덕*

*동의대학교

A Method to Process Spatial Information in Parallel Spatial DBMS

JinDeog Kim*

*Dongeui University

E-mail : jdk@deu.ac.kr

요 약

최근 공간 정보는 생산 되는 양과 데이터의 생성 빈도 및 다양성으로 인해 기존의 공간 데이터베이스 시스템에서 처리하기 어렵다. 그래서 공간 정보는 빅데이터와 연계에 관한 시도가 활발히 진행되고 있다. 그러나 효율적인 단일할당, 다중할당 색인기반 공간 연산에 대한 연구는 거의 없다.

이 논문에서는 공간 연산 중 비용이 매우 큰 공간 조인을 빅데이터 시스템에서 처리하기 위한 고려요소를 제시하고자 한다. 구체적으로 맵리듀스 시스템의 태스크 할당을 위한 단일 할당 공간 색인 방안을 설명하고, 불균일 분포가 심한 공간 정보의 특성을 고려한 부하 균등화 시 고려 요소를 제시하고자 한다. 맵리듀스와 같은 병렬 공간 데이터베이스 시스템에서의 두 가지 문제인 데이터 불균일 분포 문제와 경계 겹침 색인의 문제와의 연관성을 기술한다.

키워드

공간정보, 병렬처리, 맵리듀스, 부하 균등화

Ⅰ. 서 론

공간 정보 시스템은 공간적인 위치를 좌표 정보로 표현하는 지도정보와 그 형태와 의미를 부여하고 보완하는 비도형 속성정보를 데이터베이스의 관리기능과 연계하여 의미있는 정보를 저장, 추출, 관리, 분석하여 사용자의 의사 결정을 지원하는 시스템이다. 그러나 최근 모바일 기기, 센서 등의 보급으로 이전보다 매우 다양(variety)하고, 빈번(velocity)하면서, 대량(volume)으로 공간 정보가 생성되고 있어 기존 공간 정보 시스템에 처리하기 어려워 점이 있다.

그래서 최근 공간 정보를 빅데이터 시스템에서 처리하고자 하는 시도가 증가하고 있다. 이러한 시도록 기존의 단순 지도기반 서비스에서 공간과 다양한 서비스의 융합체인 스마트 공간 서비스가 도래하고 있다. 공간 빅데이터는 공간 분석 시각화와 함께 그 응용 분야가 점점 더 다양화 되고 있다.

이와 함께 기존 빅데이터 시스템을 확장한 공간 빅데이터 저장 관리 시스템도 다양하게 제시되고 있으며[1,2], 공간 연산자와 위치기반 인덱스를 지원하고 있다.

한편, OGC[3]에 제시된 다양한 공간 연산자 중 공간 조인은 그 비용이 매우 큰 연산자로서 OpenStreetMap[4]과 같은 데이터 집합을 처리하기 위해서는 병렬처리가 불가피하다.

그러나 현재까지 공간 빅데이터시스템에서 비용에 매우 큰 공간 조인 등과 같은 공간 연산에 대한 효율적인 색인 기법 및 부하 균등화 방안 등은 제시되지 않고 있다.

따라서 이 논문에서는 공간 조인을 빅데이터 시스템에서 처리하기 위한 고려요소를 제시하고자 한다. 특히 불균일 분포에 데이터 집합에 대한 부하 균등화를 고려한 공간 색인시 고려요소를 제시하고, 공간을 차지하고 있는 다영역에 포함되는 객체의 처리시 고려 요소를 제시하고자 한다. 그리고 이를 원활히 하기 위한 관리 테이블에 대해 설명하고자 한다.

이 논문의 구성은 다음과 같다. 2장에서는 병렬 공간 데이터베이스시스템에 관한 관련연구에 대해 기술하고, 3장에서는 이 논문에서는 제시하는 공간 조인을 빅데이터 시스템에서 처리하는 방안을 자세히 설명하고자 한다. 그리고 4장에서 결론을 맺는다.

II. 관련 연구

하둡 맵 리듀스 시스템[1]을 확장하여 공간 자료형과 공간 연산을 지원하는 공간 빅데이터 시스템 다양한 연구[5,6]가 있다. 이들 시스템은 OGC에서 제안하는 공간 데이터 타입과 공간 관계연산자와 분석연산자를 제공하고 있다.

Hadoop-GIS[6]는 공간 조인을 위한 다중 할당 경계 객체 처리 색인인 R*-tree를 이용하고, 중복 결과를 제거하는 병렬 공간DBMS를 제안하였다.

Spatial Hadoop[5]은 2단계 공간 색인을 이용한 영역질의, k-NN 질의, 공간 조인을 지원한다. 공간 색인은 그리드 파일과 R-트리와 R+-트리를 이용한다. 그리그 파일과 R+-트리는 다중할당이며, R-tree는 경계 겹침 색인이다.

Spatial Tajo[7]는 아파치 Tajo에 오픈 소스 라이브러리를 이용해 공간 연산을 지원하는 시스템이다. 빠른 공간 연산 처리를 위해 R-tree공간 색인을 이용하였다. 그리고 파티셔닝을 위해 STR을 이용하였다.

관련연구 [8]에서는 공간 빅데이터 시스템에서 보조 색인의 용량도 커짐을 고려하여 색인의 갱신 갱신 횟수를 줄이는 방안을 제시하였다. 갱신 요청 데이터를 그룹화하고 해시 테이블을 이용하여 갱신 빈도를 줄이고 있다.

III. 병렬 공간 조인 고려 사항

3.1 단일 할당 다중 조인 공간 색인

공간 조인을 위한 기존 연구는 대부분 다중할당 단일 조인 방식을 택하지만 연산횟수 증가와 연산 결과의 중복이라는 단점이 있다. 따라서 공간 색인으로서 맵 리듀스 시스템의 파티셔닝과 연산 단계의 축소가 가능한 단일할당 다중 조인 방식의 그리드 파일이 사용가능하다. 이때 다단계 확장MBR과 대응그리드 설정기법이 필요하다.

3.2 부하 균등화

부하 균등화를 위해서는 각 그리드 셀의 작업량을 사전 측정한 빈도테이블을 토대로 추정한다.

$$T_i = \sum_{j=1}^{k} |R_i| * |S_j|$$

Ti : workload of a task i (i = 1 ~ n)
k : the number of corresponding grids
|Ri| : the number of objects of grid cell Ri
|Sj| : the number of objects of grid cell Sj
(where Sj are the corresponding grid of Ri)

오름차순으로 정렬된 태스크 리스트를 모두 할당할 때가지 라운드 로빈 방식으로 진행한다.

비록 정적 태스크 할당으로 태스크 부하가 일정해도 공간 연산의 특성상 연산 시간의 모호성이 항상 존재하고, 빅데이터 시스템의 특성상 부노드의 성능을 실시간 적용해야 하므로 준동적 태스크 할당이 필요하다.

준동적 태스크 할당은 태스크 중 일정 부분을 정적으로 인접성을 고려하여 태스크를 할당한 뒤, 나머지 부분은 동적으로 크기를 고려하여 태스크를 할당한다. 이는 인접성을 고려해 디스크 검색의 중복을 막고 부노드의 성능을 고려한 동적으로 부하 평준화를 수행할 수 있는 방법이다. 정적 할당의 비율은 데이터의 분포 테이블을 활용하여 불균일 분포 문제를 해결한다. 이 때 인접도의 기준은 Hilbert Curve의 차이값이다.

IV. 결론

이 논문에서는 최근 활성화되고 있는 공간 빅데이터 시스템에서 비용이 큰 공간 조인과정의 고려 요소를 제시하였다.

맵 리듀스 시스템에 적용하기 위해서는 파티셔닝이 필요하며, 이를 위해 영역 비겹침 색인이 필요하고, 분산 처리의 중복 제거라는 불필요한 과정을 생략하기 위해서는 단일 할당 다중 조인공간 색인이 바람직하다.

그리고 빅데이터의 시스템 확장성(Scale Out)과 공간 데이터의 분균일성을 고려할 때 빈도 정보를 활용한 정적 태스크 할당과 부노드들의 자원의 차이를 감안한 준동적 태스크 할당이 요구됨을 알 수 있다.

참고문헌

[1] Hadoop-GIS : http://hadoopgis.org
[2] MongoDB : http://www.mongodb.org
[3] OGC : http://opengeospatial.org
[4] OpenSteetMap: http://openstreetmap.org
[5] A. Eldawy, M. F. Mokbel, “A Demonstration of SpatialHadoop: An Efficient MapReduce Framework for Spatial Data”, Proc. of VLDB Endowment, Vol.6, No. 12, 2013
[6] A Aji, F. Wang, H. Vo, R. Lee, Q. Liu, X. Zhang, J. Saltz, “Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce, Proc. of VLDB Endowment, Vol.6, No.11, 2013
[7] 조현구, 유기현, 양평우, 이연식, 남광우, “Spatial Tajo : 공간 빅 데이터 웨어하우스 시스템의 설계 및 구현”, 한국정보과학회 동계학술대회, 2015
[8] 최용권, “위치 정보 데이터가 포함된 빅 데이터 환경에서 시공간 인덱스 갱신 효율을 위한 MapReduce 확장”, 인하대학교 대학원 석사, 2012