공간데이터베이스관리시스템의 성능 평가

이 정민° , 이 기준
부산대학교 전자계산학과

Benchmarking Spatial DBMSs

Jeong-Min Lee and Ki-Joune Li
Department of Computer Science, Pusan National University
e-mail : {jmlee, lik}@spatios.cs.pusan.ac.kr

요약

공간데이터베이스관리시스템은 공간 자료를 처리하기 위한 여러 응용에서 사용되는데, 기존의 비공간 자료를 처리하던 데이터베이스관리시스템과는 다른 기능 및 성능이 요구된다. 공간 데이터베이스시스템의 기능과 성능이 사용자의 요구를 만족하는지를 알기 위해서 시스템에 대한 성능 평가를 행하는데 이를 벤치마크라 한다. 본 논문에서는 공간데이터베이스시스템에서 요구하는 기능적 요구사항이 무엇인지를 알아보고 이들의 성능을 평가하기 위한 환경을 제안한다. 시스템의 성능에 영향을 끼치는 공간 자료나 공간 질의의 요인들을 살펴보고 이를 기반으로 테스트용 공간 데이터베이스와 공간 질의를 생성한다. 또한 성능 평가의 척도로 질의 수행 시간, 디스크 참조 횟수, 메모리 요구량 등을 정의하였다. 특히 본 논문은 공간데이터베이스관리시스템의 벤치마크를 위해 인위적으로 합성된(synthetic) 자료나 질의를 생성하는데 관한 기술을 하고 있는데 궁극적으로 벤치마크를 위한 자료나 질의를 생성하는 환경을 개발하는 것이 목적이다.

1. 개요

공간데이터베이스관리시스템은 기존의 비공간 자료를 다루던 시스템과는 다른 기능과 성능이 요구된다. 공간데이터베이스관리시스템의 기능과 성능이 응용에서 요구하는 성능을 만족하는지를 알기 위해서는 시스템에 대한 성능 평가를 행해야 하는데 이를 벤치마크라 한다. 벤치마크란 시스템의 성능을 평가하기 위한 전체적인 실험 환경을 통틀어 일컫는데 이는 시스템의 성능을 평가하는 목적 뿐만아니라 시스템의 취약 부분을 찾아내어 이를 보완함으로써 전체적인 시스템 성능 향상을 도모하기 위해서도 사용되고 있다.

기존의 데이터베이스관리시스템의 성능을 평가하기위한 여러 가지 벤치마크들이 제안되었다. 관계형 데이터베이스관리시스템[1]이나 트랜잭션 처리 시스템[2,3], 공학용 데이터베이스관리시스템[4] 그리고 객체지향 데이터베이스관리시스템[5]을 평가하기 위한 것 등이 있다. 그러나 이러한 벤치마크들은 공간 객체들로 구성된 데이터베이스를 구축하거나 공간 색인 기법의 사용 또는 공간 질의를 수행하는 작업들을 거의 고려하고 있지 않아서 공간데이터베이스관리시스템을 위한 벤치마크로는 쓰일 수 없다.

이에 비해 Sequoia 2000벤치마크[6]는 공간데이터베이스시스템을 평가하기 위한 환경을 제안했다. 래스터이미지나 점, 선, 면과 같은 공간 객체를 포함하는 데이터베이스를 구축하고 여기에 여러 가지 형태의 공간 질의를 수행하여 질의 처리 시간을 측정하였다. 성능 척도로는 전체 질의 처리 시간 대 시스템의 가격비를 정의하였다. 여기에서 정의된 공간 데이터베이스는 실제 데이터를 이용하여 데이터의 모양, 데이터베이스의 크기에 따라 몇가지로 구분하였고 공간 질의는 몇가지 간단한 형태의 질의로 구성되었다. 그러나 Sequoia 2000은 공간데이터베이스시스템의 제한된 기능 및 성능을 측정하는데 국한되어 있어, 일반적인 공간 데이터베이스관리시스템의 벤치마크로는 부적절하다. 또한 성능에 영향을 끼치는 더 많은 요소들에 의해 다양화될 수 있고 성능 평가의 목적으로 사용될 수 있는 척도도 여러 가지가 있을 수 있다.

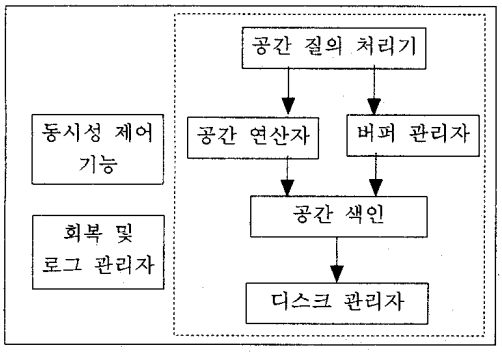
본 논문에서는 벤치마크를 위하여 실제 데이터가 아니라 인위적으로 합성된 다양한 종류의 공간 데이터나 공간 질의를 생성하고자 한다. 실제 데이터는 공간 자료의 부분적 특성만을 반영하게 되므로 다양한 실험을 어렵게 만드는 반면, 인위적으로 합성된 자료는 여러 가지 요인에 의해 다양하게 생성할 수 있다. 여기서는 테스트용 데이터나 질의 설계시 성능에 영향을 끼칠 수 있는 요소들에 대하여 알아보고 이들을 바탕으로 공간 자료나 공간 질의를 생성하는 방법을 설명한다. 그리고 의미있는 대표적인 데이터와 질의의 조합을 구성하고 여러 가지 성능 평가 척도를 제시한다. 그리하여 궁극적으로 벤치마크를 위한 툴을 개발하는 것이 목적이다.

본 논문의 구성은 다음과 같다. 2장에서는 공간데이타베이스관리시스템의 기능적 요구 사항과 함께 논문에서 제시하는 벤치마크의 목적과 그 범위에 대하여 논한다. 벤치마크 데이터베이스를 구현할 때 성능에 영향을 끼칠 수 있는 공간 자료의 요인들은 3장에서 제시하고 질의 구현시의 요인들은 4장에서 제안한다. 5장에서는 이러한 요인들을 바탕으로 의미있는 공간 자료와 공

간 질의의 조합을 구성한다. 그리고 성능을 평가할 때 어떤 기준으로 할 것인지 성능 척도에 대해서는 6장에서 알아보고 마지막으로 7장에서는 본 논문의 결론을 짓고 향후 연구 방향에 대하여 기술한다.

2. 요구사항 및 목적

공간데이터베이스관리시스템은 응용 시스템들의 기능과 성능을 좌우하는 아주 기본적인 기능을 수행한다. 공간데이터베이스관리시스템의 일반적인 시스템 구조를 살펴보면 [그림 1]과 같다.



[그림 1] 공간데이터베이스관리시스템의 구조

공간 질의 처리기는 사용자가 던지는 다양한 공간적 질의를 여러 개의 공간 연산자로 나누고 질의 최적화를 통해 질의 수행 계획을 세운다. 공간 연산자는 공간 질의를 수행하기 위한 기본적인 공간 연산을 수행하는데 일반적으로 공간 색인 기능과 강하게 결합되어 있다. 공간 색인을 통해 공간 자료에 접근하는 것은 여러 번의 디스크 참조를 요구하게 되고 이는 수행 속도를 떨어뜨리는 요인이 되므로 버퍼 관리자는 자료에 접근할 때 디스크 참조 횟수를 줄여 성능을 향상시킨다. 공간 자료는 일반적으로 자료를 디스크에 저장하는 과정이 복잡할 뿐만 아니라 용량 또한 매우 크다. 따라서 공간 자료에 대한 검색 속도를 빠르게 하기 위해서 디스크에 저장된 자료에 대한 색인 정보를 따로 두어 이를 이용한다. 효율적인 공간 색인 기법은 불필요한 검색을 줄여주고 검색 속도가 빨라야 하며 색인 파일의 크기가 너무 커도 곤란하다. 디스크 관리자는 실제로 디스크에 공간 자료를 저장하고 이를 관리하는 부분이다. 디스크관리자가 데이터 클러스터링 기법에 의해 검색시 같이 검색될 확률이 많은 데이터들을 같은 페이지내에 저장함으로써 주어진 질의에 대한 디스크 참조 횟수를 감소시킬 수 있다. 동시성 제어 기능은 다수의 사용자가 시스템을 동시에 사용하고자 할 때 이를 제어하는 기능이다. 동시에 서로 다른 사용자가 동일한 데이터베이스를 사용하려고 할 때 자료의 불일치성이 생기지 않도록 작업들을 스케줄링한다. 회복 및 로그 관리자는 시스템의 기능 장애로 작업이 실패하게 되는 경우 데이터베이스에 발생하게 되는 자료의 불일치성이나 작업 오류 같은 것을 바르게 정정할 수 있는 기능을 말한다.

공간데이터베이스관리시스템을 다양한 응용에서 사용하려고 할 때 시스템의 이러한 기능들이 응용에서 필요로하는 성능적 요구 사항을 만족하는지를 알아야 한다. 본 논문에서는 공간데이터베이스관리시스템에서 사용되는 공간 데이터나 공간 질의들이 위에서 보인 여러 가지 모듈들의 성능에 영향을 미치는 요인에 대해 알아보고 이를 기준으로 벤치마크를 위한 데이터와 질의들을 생성한다. 본 논문의 목적은 인위적으로 합성된 데이터와 질의들로 테스트를 할 수 있도록 벤치마크를 위한 실험환경을 만드는 것이다. 동시성 제어 기능과 회복 및 로그 관리자의 기능에 대한 테스트는 제외한다.

3. 성능에 영향을 주는 데이터베이스 요인들

본 장에서는 벤치마크용 공간 데이터베이스를 설계할 때 성능에 영향을 미칠 수 있는 여러 가지 요인들에 대해 고려해 본다. 그리고 이러한 요인들을 측정할 수 있는 기준을 정의한다.

3.1 공간 자료의 형태

공간 자료는 응용의 요구에 따라 다양한 형태를 가진다. 이차원 이상의 다차원 형태일 수도 있고 단순한 모양에서부터 복잡한 모양에 이르기까지 다양하다. 기본적인 공간 자료는 점, 선 그리고 다각형이다. 공간 자료의 형태에 따라 이를 처리하는 방법들도 다르게 나타나는데 차원이 높아지고 모양이 복잡해질수록 더욱 복잡한 처리 기법이 요구되고 이는 성능에 영향을 미치게 된다.

3.2 데이터베이스의 크기

테스트하는 자료의 양을 어느 정도로 하느냐에 따라 시스템의 성능이 다르게 나타날 수 있다. 벤치마크에서는 보통 Small, Medium, Large로 데이터베이스를 구분해서 성능을 테스트한다.

3.3 공간 자료의 상대적 크기

처리하고자 하는 공간 자료의 상대적 크기는 공간 자료가 전체 도면 또는 작업 공간에서 차지하는 크기를 말한다. 공간 자료의 크기가 크면 상대적으로 공간 자료들이 서로 겹칠 확률이 높을 것이고 이는 공간 색인 기법의 효율에 영향을 미친다. 실제로 공간 자료의 크기가 도면 또는 작업 공간의 크기에 비해 차지하는 비율은 응용마다 다양하게 나타난다. 그러므로 다양한 크기의 공간 자료에 대한 테스트가 요구된다.

공간 자료의 상대적 크기는 전체 공간에서 공간 자료가 차지하는 크기로 공간 자료의 개수를 N이라 할 때, 크기는 $[0, 1/\sqrt{N}]$의 범위에 있도록 한다. 크기가 $1/\sqrt{N}$ 이상일 경우엔 전체 공간 자료가 차지하는 면적 $S > ((1/\sqrt{N} \times 1/\sqrt{N}) \times N) = 1$이 되어서 주어진 전체 공간의 넓이를 초과할 수 있다.

3.4 공간 자료의 밀도

밀도란 단위 영역내에 존재하는 공간 자료의 수로 정의할 수 있는데 이는 공간 자료가 공간상에서 얼마나 밀집되어 존재하는지를 나타내는 척도이다. 동일한 크기의 영역에서 존재하는 공간 자료의 수가 많으면 밀도가 높고 그렇지 않으면 밀도가 낮다.

공간 자료의 밀도는 다음과 같이 정의된다.

$$D(A) = n/S(A)$$

단, n은 A지역에 있는 공간 객체의 수이고 $S(A)$는 A지역의 넓이이다.

3.5 공간 자료의 분포

모든 현상은 어느 특별한 경우에 집중되어 일어난다는 Zifp's Law를 공간 자료의 분포에 적용하면 공간 자료는 주어진 공간에 균일하게 분포되어 있기보다는 어느 특정한 지역에 집중되어 분포되어 있다. 실제로 공간 자료들은 특정 부분에 밀집되어 존

재하는 경우가 대부분이고 그 밀집된 정도도 또한 다르다.

공간 자료의 분포는 밀도를 이용해 정의할 수 있다. 공간 자료의 집중된 정도는 공간 상에서 단위 공간당 공간 자료의 밀도가 거의 비슷하면 균등 분포에 가깝고 밀도차가 클수록 집중된 정도가 커짐을 알 수 있다. 그러므로 공간에 분포해 있는 공간 자료의 집합 S의 밀도 분산를 통해서 집중된 정도를 측정할 수 있다. 본 실험에서는 임의의 단위 원을 떨어뜨려 밀도를 계산하고 그 분포를 통해 밀도 분산을 다음과 같이 정의한다.

$$Var_{density}(S) = \left(\sum_{i=0} (n_e - n_i)/(m-1) \right)^{1/2}$$

단, n_i는 넓이가 1인 i번째 임의의 원에 포함되는 객체의 수이고, n_e는 임의의 원에 포함되는 공간 객체수의 평균, m은 전체 공간에 떨어뜨리는 원의 갯수이다. 여기서, 우리는 다음과 같은 것을 알 수 있다.

i, 만약, $Var_{density}(S)$= 0이면 균등 분포이다.
ii, 만약, 두 공간 객체 집합 *S1*, *S2*가
$Var_{density}(S1) < Var_{density}(S2)$이면 *S2*가 더욱 skewed되었다.

3.6 공간 자료의 복잡도

선이나 다각형과 같은 공간 자료는 다양한 모양을 가질 수 있는데 공간 자료의 형태가 복잡할수록 복잡한 연산이나 저장, 관리 방법을 요구하게 된다. 다각형의 복잡도는 [8]에서 정의한 복잡도를 사용하는데, 이는 다각형을 구성하는 점의 개수, 점 사이의 거리, 점의 진동의 횟수나 크기, 볼록다각형(Convex Polygon)으로부터 벗어난 정도에 의해 계산된다.

4. 성능에 영향을 주는 질의 요인들

질의는 질의 처리기를 통해 공간 연산자로 나뉘어져서 수행되고 이 때의 성능은 질의 처리기나 공간 연산자 뿐만 아니라 공간 색인, 버퍼 관리자등과도 밀접한 관련이 있다. 여기서는 성능 평가에 영향을 미치는 공간 질의의 요인들과 이들의 측정하는 기준을 제시한다.

4.1 질의의 종류

공간데이터베이스관리시스템에서는 우선 데이터베이스를 만들거나 자료를 삽입, 삭제, 수정하는 기능이 기본적으로 제공된다. 이는 공간 자료를 공간 데이터베이스에 저장, 관리하는 기능들로 공간 색인 파일과 데이터 파일을 생성한다.

공간 질의로는 최소거리, Buffer Zone, 넓이, 연결성에 관한 질의나 Up-stream/Down-stream, 겹침, 포함, 인접 관계 등이 있다. 단순히 공간 자료의 특성값을 구하는 질의일 경우는 CPU 의존적이지만, 대부분의 질의는 I/O 의존적이면서 복잡한 작업이 요구된다. 공간 질의를 수행하기 위한 공간 연산자는 공간 색인을 통하여 공간 자료를 접근하므로 공간 색인의 성능과 공간 연산 알고리즘, 버퍼 관리자의 성능에 의존적이다.

공간 죠인은 두 개 이상의 도면을 두고 공간 자료의 위치에 기반하여 공간상에서 서로 겹치는 공간 자료를들 검출한다. 공간 죠인은 대량의 공간 자료를 다루고 많은 비교 연산을 수행하여 처리시간이 많이 든다.

4.2 공간 질의의 분포

공간 자료와 마찬가지로 공간 질의도 특정 지역에 집중되어 일어난다. 공간 자료의 밀도가 높은 지역에서는 검색할 객체가 다양하고 정보량도 많아서 자주 검색 지역에 포함될 가능성이 높은 반면 밀도가 낮은 지역은 검색의 대상이 될 만한 공간 자료의 수가 적으므로 실제로 검색의 범위에 포함될 가능성이 낮다. 그러므로 전체적으로 질의는 공간 자료의 밀도가 높은 지역에 공간적으로 집중되어 나타나게 된다. 단위 영역당 관련된 질의의 수를 공간 질의의 밀도라고 한다면 공간 질의의 분포는 공간 자료에서와 같이 밀도를 이용하여 측정할 수 있다.

4.3 공간 질의의 상대적 크기

공간 질의의 상대적 크기는 주어진 공간에서 질의 영역이 차지하는 비율이다. 실제로 질의 영역은 주어진 공간에서 너무 크거나 너무 작으면 비현실적이다. 공간 질의의 상대적 크기는 공간 자료의 상대적 크기와 같이 측정한다.

4.4 공간 질의의 시공간적 집중성

일련의 질의들은 서로 아무런 관련없이 주어지는 것이 아니라 특정 목적을 위해 그 이전 질의와 그 결과에 관련하여 주어진다. 이전 질의에서 검색된 공간이 다음 질의에 다시 검색될 가능성이 높은데 이는 공간 질의가 시간적 순서에 의한 공간적 집중성을 가지게 된다는 것을 의미한다. 즉 연속적으로 주어지는 질의 영역이 공간적으로 서로 가깝게 위치할 확률이 높다는 것이다.

공간 질의의 시공간적 집중성은 다음과 같이 질의거리의 비율로 정의할 수 있다. 질의 영역이, Q_0, Q_1, Q_2, ... Q_{n-1}와 같이 주어졌을 때 질의거리의 비율 $Dist_{query}$는 다음과 같이 정의된다.

$$Dist_{query} = \sum_{i=1} (Q_i - Q_{i-1})/(n \times d)$$

단, d는 공간 상의 최대거리이다. 즉 질의들이 시간적 공간적으로 집중되어 있을 때 거리가 작아짐을 알 수 있다. $Dist_{query}$는 [0, 1]의 값을 가질수 있는데 실제로 0.5일 때 거의 무작위로 떨어지는 것과 같다.

5. 테스트용 데이터베이스 및 질의 설계

앞에서 논의한 성능에 영향을 미치는 여러 가지 요인들을 기준로 하여 벤치마크에 이용될 공간 데이터베이스와 공간 질의를 생성한다. 본 장에서는 공간 데이터와 질의의 조합으로 벤치마크 환경을 구성하는데 테스트에 적당한 각 요인들의 범위를 함께 나타내었다.

5.1 데이터베이스 구축 및 관리

DB Build : 주어진 공간 자료를 위한 색인 파일과 데이터 파일을 구성한다.					
질의	데이터				
형태	DB 크기	형태	공간자료 크기	분포	복잡도
삽입 삭제 수정	Small Medium Large	점 선 다각형	0.001 0.005 0.01	0.0 0.3 0.6	0.0 0.3 0.6

5.2 공간 질의

nearest : 임의의 점을 주고 가장 거리가 가까운 공간 자료를 찾는다			
질의	데이터		
형태	형태	크기	분포
점	점	0	0.0 0.3 0.6

containment : 질의 영역에 포함되는 공간 자료를 찾는다.							
질의				데이터			
형태	질의 크기	분포	시공간적 집중도	형태	크기	분포	복잡도
사각형	0.05 0.2 0.3	0.0 0.3 0.6	0.1 0.3 0.5	다각형	0.001 0.005 0.01	0.0 0.3 0.6	0.0 0.3 0.6

network trace : 임의의 선과 방향을 주면 주어진 방향으로 연결된 선을 찾는다.			
질의	데이터		
형태	형태	분포	복잡도
선, 탐색 방향	선	0.0 0.3 0.6	0.0 0.3 0.6

5.3 공간 죠인

Spatial Join : 두 개의 도면을 두고 서로 겹치는 공간 자료를 찾는다.				
질의	데이터			
형태	형태	크기	분포	복잡도
intersect (선) overlap (면)	선 면	0.01 0.1	0.0 0.3 0.6	0.0 0.3 0.6

6. 성능 평가 척도

본 장에서는 공간데이터베이스관리시스템의 성능을 측정하기 위한 성능 평가 척도에 대해서 알아본다. 성능 평가 척도란 시스템의 성능을 측정할 때 성능의 좋고 나쁨을 결정하는 기준이 되는 것을 말한다.

6.1 질의 수행 시간

질의 수행 시간 시간이란 사용자가 시스템에 어떤 작업을 지시한 순간부터 작업의 결과를 돌려받기까지 걸리는 시간을 말한다. 사용자가 지시한 작업의 종류에 따라 수행 시간을 측정하는데 먼저 공간 데이터베이스구축 및 관리 작업으로 공간 자료를 삽입, 삭제 그리고 수정하는 작업에 대해서 처리 시간을 측정한다. 공간 죠인을 포함한 공간 질의는 위에서 보인 주어진 데이터에 대해 각각의 질의 수행시간을 측정한다.

6.2 디스크 참조 횟수

질의 처리 시간에서 디스크 참조 시간을 포함하고 있지만 여기서는 디스크 참조 횟수를 따로 평가하려고 한다. 시스템을 구성하는 어떤 부분이 취약한지를 알아내고자 할 때 응답 시간만을 성능 척도로 했을 경우 알아내기 힘들다. 질의를 처리하면서 드는 CPU비용과 디스크 참조 비용을 구분함으로써 성능을 분석하는데 도움이 될 수 있다.

일반적으로 대량의 데이터베이스를 다루는 시스템에서는 CPU연산에 드는 비용보다 디스크를 참조하는데 드는 비용을 더욱 중요하게 여긴다. CPU의 속도는 디스크를 움직이는 속도에 비해 훨씬 빨라서 주로 질의 처리에 드는 시간은 디스크를 참조하면서 걸리는 시간에 의존한다. 버퍼 관리자나 공간 색인 또는 클러스터링 기법이 얼마나 좋은 성능을 나타내는가는 디스크 참조를 얼마나 적게 하면서 질의를 처리하는지로 나타난다.

6.3 메모리 요구량

응답 시간과 디스크 참조 횟수만으로는 객관적인 성능을 평가하는 기준이 될 수 없다. 왜냐하면 메모리 요구량이 모두 동일한 환경에서의 비교가 아니기 때문이다. 메모리 요구량은 구축한 데이터베이스를 유지하는데 드는 비용으로 색인 파일과 공간 자료 파일을 모두 포함한다. 메모리 요구량은 데이터베이스를 얼마나 효율적으로 구축하였는가를 측정하는 것이다.

7. 결론 및 향후 연구 방향

공간데이터베이스관리시스템의 성능을 평가하기 위해 인위적인 공간 데이터와 공간 질의를 생성하고자 할 때 고려해야 할 사항에 대해 알아보았다. 그리고 이렇게 생성된 데이터와 질의로 성능을 측정할 때 평가 기준이 되는 척도에 대해서도 알아보았다. 앞으로 프랙탈을 이용하여 데이터나 질의의 분포[9]도 고려할 예정이며 본 논문에서 제시한 질의와 데이터들을 생성할 수 있는 벤치마크 환경은 현재 개발중이며 후에 공개할 예정이다.

참고 문헌

[1] Bitton, D., et al., "Benchmarking Database Systems: A Systematic Approach," Proc. VLDB Conference, Florence, Italy, Nov. 1983

[2] Anon et al., "A Measure of Transaction Processing Power," Datamation, 1985

[3] Scott T. L. and Daniel A. , "A Modeling Study of the TPC-C Benchmark," Proc. ACM SIGMOD, 1993

[4] Cattell, R.G.G., and Skeen, J. , "Engineering database benchmark," ACM Transactions on Database Systems, 1991

[5] Carey M. J., DeWitt D. J. and Naughton J. F., "The 007 Benchmark," ACM SIGMOD, 1993

[6] Stonebraker M., Frew J., Gradels K. and Meredith J., "The Sequoia 2000 Storage Benchmark," Proc. ACM SIGMOD, 1993

[7] Kim M. S., Shin Y. S., Cho M. J. and Li K. J., "A Comparative Study of Spatial Access methods," Proc. ACM-GIS 1995

[8] Brinkhoff T., Kriegel H. P., Schneider R. and Braun A., "Measuring the complexity of polygonal objects," Proc. ACM-GIS 1995

[9] Belussi A., Faloutsos C., "Estimating the Selectivity of Spatial Queries Using the 'Correlation' Fractal Dimension," Proc. VLDB Conference, Zurich, Switzerland, 1995