

균등분포로 변환을 이용한 공간 색인방법

문 정욱^o, 김 동현, 이 기준
부산대학교 전자계산학과

{jwmoon, dhkim}@quantos.cs.pusan.ac.kr, lik@hyowon.pusan.ac.kr

Spatial Indexing Methods by Transformation to Uniform Distribution

Jung-Wook Moon^o, Tong-Hyon Kim, and Ki-Joune Li
Department of Computer Science, Pusan National University

요 약

대부분의 공간색인 방법들은 공간객체의 분포가 균일하지 않다는 가정 하에 만들어졌다. 이 가정은 실제의 응용분야에 일반적으로 적용된다. 반면에 균등한 공간객체의 분포를 위하여 만들어진 공간색인 방법들도 몇 가지 있다. 예를 들어, 고정 격자 방식의 공간색인 방법은 균등한 공간객체에 대한 색인 방법으로 매우 좋은 성능을 보이며, 색인과정이 매우 단순하다는 장점이 있다. 본 논문에서는 비균일적인 공간객체의 분포를 균일 분포로 변화한 다음 고정 격자 방식의 공간색인을 적용하는 방법을 제안한다. 이 방법에 따르면, 공간효율성이 매우 증가하고, 빈 공간이 줄어들어 드는 등 공간색인의 성능이 다른 색인방법에 비하여 매우 증가하며, 색인과정이 단순하여 지는 장점을 가지고 있다.

1. 서 론

지금까지 제안된 대부분의 공간색인 방법은 비균등 공간객체의 분포에도 좋은 성능을 보일 수 있도록 만들어진 것이다. 예를 들어, R-tree 계열의 공간색인 방법[1]은 공간객체의 분포에 상관없이 좋은 성능을 보일 수 있도록 만들어진 색인 방법이다. 반면에 격자 파일이나 고정 격자 파일과 같은 공간 색인 방법[2]은 비균일 공간객체에 대하여서는 매우 떨어지는 성능을 보이지만, 균일한 공간객체의 분포에 대하여서는 매우 좋은 성능을 보여준다. 또한 이들 공간색인 방법들은 단순하여 구현하고 관리하기가 매우 쉽다는 장점도 가지고 있다.

그러나, 이와 같은 장점에도 불구하고, 대부분의 응용분야에서 주어지는 실제 공간객체의 분포는 비균일 분포이어서, 이들 방법은 사용되기 곤란하다. 본 논문에서는 이들 방법, 특히 고정 격자 방법의 장점을 활용하기 위한 방법을 제안한다. 즉, 주어진 비균일적인 공간객체를 균일분포로 전환하여, 고정 격자 방법을 적용한다. 원래 고정 격자 방법에 의한 공간색인은 여과단계를 위하여 색인 구조에 대한 디스크 접근이 불필요하고 단지 정제단계를 위한 데이터 블록에 대한 디스크 접근만이 요구되어 매우 성능이 뛰어나다. 따라서, 본 논문에서 제안한 변환방법을 이용하면, 비균일 공간객체에 대하여서도 고정 격자 방법의 장점을 그대로 살릴 수 있다.

본 논문에서는 먼저, 2장에서 고정 격자 방법의 문제점과 장점, 그리고 본 논문의 동기에 대하여 간단하게 살펴보고, 3장에서는 균등분포로 변환하는 방법을 제시한다. 4장에서는 고정 격자 방법과 균등분포로 변환하는 방법을 이용

한 색인방법을 제시하고 5장에서는 본 논문의 결론을 맺는다.

2. 고정 격자 공간색인방법의 장점과 단점

공간색인의 성능은 공간객체의 분포에 많은 영향을 받는다. 또한 대부분의 응용분야에서 주어지는 공간객체의 분포는 비균일적이기 때문에, 공간색인은 비균일적인 공간객체의 분포에 대하여서도 좋은 성능을 보여야만 한다. [3]의 연구에서는 일반적인 공간객체의 분포는 비균일적이며, 대략 1.6 프랙탈 차원에 해당한다고 하였다. 따라서, 현재 사용되고 있는 대부분의 공간색인방법은 비균일적인 공간객체에 대하여서도 균일분포의 공간객체와 비슷한 안정적인 성능을 보여주고 있다. 예를 들어, R-tree 계열의 공간색인방법은 공간객체의 분포와 거의 무관하게 좋은 성능을 보여 주고 있다.

그러나, 공간객체의 비균일 분포의 문제를 극복하기 위하여, R-tree 계열의 방법은 비교적 복잡한 구조와 과정을 통하여 색인을 수행한다. 먼저, 색인을 위하여서는 트리구조를 가지는 자료구조를 필요로 하고, 공간질의처리를 위하여서는 이 자료구조에 대한 비교적 복잡한 처리과정이 요구된다. 또한 공간색인을 위하여서는 이 트리자료 구조의 방문을 위하여 많은 디스크 접근이 필요하다. 이러한 과정을 통하여서만 원하는 공간객체가 저장되어 있는 디스크 주소를 알 수 있다.

만일 공간객체의 분포가 균일하게 되어 있으면,

위와 같은 복잡하고 많은 시간이 요구되는 과정은 필요 없게 된다. 즉, 아래의 그림 1에서와 같이 하나의 격자는 하나의 디스크 블록에 대응되므로, 질의영역과 겹치는 격자에 해당하는 디스크 블록을 읽으면 되므로 다른 공간색인 방법과 같이 색인과정의 디스크 접근이 불필요하다. 예를 들어, R-tree의 디렉토리 노드에 해당하는 디스크 블록을 읽는 시간이 불필요하고, 아래의 그림에서와 같이, 단순히 5개의 데이터 블록만 읽는 시간으로 질의처리가 가능하다.

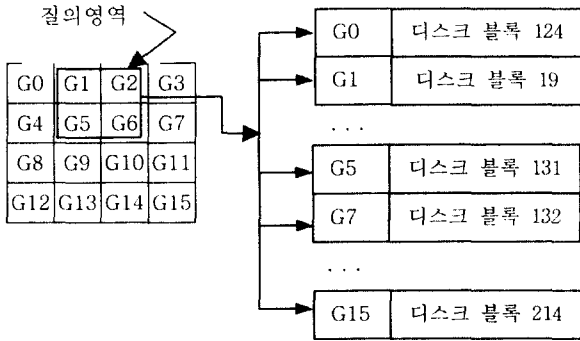


그림 1. 고정 격자를 이용한 공간색인

그러나, 고정 격자 방법의 가장 큰 단점은 사상공간(Dead Space)으로 인한 저장공간효율성의 저하이다. 앞에서 언급한 바와 같이, 공간객체의 분포는 비균일적이다. 이것은 각 격자에 속하는 공간객체의 수가 서로 다르다는 것을 의미한다. 즉, 격자의 밀도 density(G_i)가 서로 다르다는 것을 의미한다. 그런데, 하나의 격자는 하나의 디스크 블록에 대응되므로 다음과 같은 조건을 만족하도록 격자가 구성되어야 한다. 여기서 a 는 하나의 격자의 넓이를 의미한다.

$$a < B_f / d_{max} \quad -- (1)$$

여기서 B_f 는 디스크 블록의 최대 저장 객체 수(Blocking Factor)이고, $d_{max} = \text{Max}(\text{density}(G_i), i=0, n-1)$ 이다. (1)식은 B_f 는 정하여져 있으므로, 최대 밀도가 높아지면 하나의 격자의 넓이는 작아지고 결과적으로 격자의 개수와 디스크블록의 수 N_{block} 는 증가된다는 것을 의미한다. 즉, 아래의 식(2)와 같이, 저장공간의 효율성이 떨어지므로, 주어진 질의영역에 걸쳐지는 격자의 수가 많아지고, 격자에 대응되는 디스크 블록의 수가 많아져 색인과 정에 필요한 디스크 접근은 없지만 데이터 블록을 위한 디스크 접근 수가 많아진다는 것을 의미한다.

$$N_{block} = A/a \text{ 이므로 } (A \text{는 전체 영역의 넓이})$$

$$N_{block} > A d_{max} / B_f \quad -- (2)$$

만일 비균일정도, 즉 공간객체의 집중도가 커지면 d_{max} 가 커지고 결과적으로 디스크 블록의 수도 함께 많아지게 된다.

본 논문에서는 비균일적으로 분포되어 있는 공간객체

의 위치를 적절하게 변환하여 균일분포로 만드는 방법을 제안한다. 즉, 균일분포로 만들어, d_{max} 의 값을 가능한 평균밀도 d_{mean} 에 근접하게 만들어 저장공간의 효율성을 증가시키고, 결과적으로 공간질의 처리에서 요구되는 데이터 블록의 접근횟수를 줄이는 방법을 제안한다.

3. 균일분포로 변환 방법

본 논문에서 제안하는 변환방법은 두 가지이다. 첫번째 방법은 들로니 삼각분할(Delaunay Triangulation)을 이용한 탄성변형이고[4], 두 번째 방법은 [5]에서 제안된 STR 방법을 응용한 것이다.

삼각분할에 의한 탄성변형의 자세한 설명은 본 논문에서 생략하기로 한다. 자세한 설명을 위하여서는 [4]을 참조하기 바란다. 탄성변형은 주어진 제어점(Control Point)을 이용하여 임의의 점의 위치를 이동하는 기능을 제공한다. 즉, 주어진 제어점의 집합을 CP라고 하였을 때, 다음과 같은 함수로 표현된다.

$$\text{Elastic Transform}(CP, p) = p' \quad -- (3)$$

여기서, 원래 공간의 임의의 점 p 에 대하여 p' 은 변형된 공간상의 이동된 위치를 말한다.

위의 식 (3)에서 제어점의 집합을 결정하는 것이 탄성변형의 특성을 결정하는데 매우 중요한 역할을 한다. 특히, 균일분포를 얻기 위하여서는 적절한 제어점의 집합을 선정하는 것이 중요하다. 이를 위하여서, 본 논문에서는 아래의 그림 2와 같이, 주어진 점 객체를 이용한 격자의 교차점의 일부를 제어점 집합으로 정하였다.

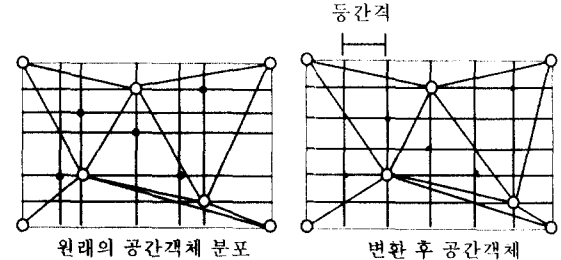


그림 2. 탄성변형을 이용한 균일분포변환

위의 그림에서 검은 점은 공간객체를 의미하고, 하얀 점은 제어점을 의미한다.

두 번째 변환방법은 STR을 이용하는 것이다. STR로 나누어지는 공간을 동일한 크기의 공간으로 변환하여, 아래의 그림 3에서와 같이 하나의 격자 내에 있는 공간객체의 위치를 변환하는 것이다.

아래 그림 3의 두 가지 방법은 모두 장단점을 가지고 있다. 우선, 공간질의 처리를 위하여서는 변형을 위한 보조적인 정보가 필요하다. 탄성변형의 경우는 제어점의 좌표가, 그리고, STR을 이용한 방법은 STR의 분할 선의 좌표가 필요하다. 그런데, 동일한 양의 보조

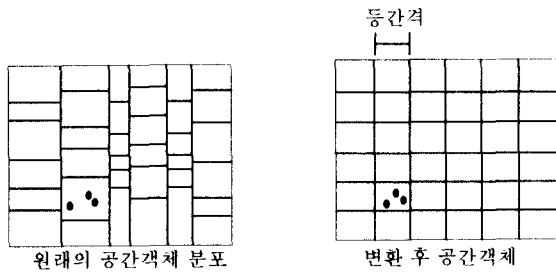


그림 3. STR을 이용한 균일분포변환

정보를 이용하여 분할을 하였을 경우, 일반적으로 STR을 이용한 방법이 보다 균일분포에 근접하는 결과를 얻을 수 있다.

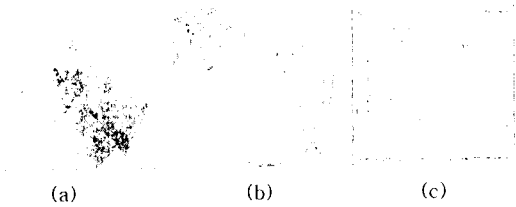


그림 4. 균일분포로 변형 예

위의 그림 4는 균일분포로 변형한 예이다. 4(a)는 미국 몽고메리지역의 데이터이고, 4(b)는 탄성변형을 이용하여 변형을 한 것이고, 4(c)는 STR을 이용하여 변형을 수행한 것이다. 위의 그림에서와 같이, STR을 이용한 것이 보다 균일분포에 가깝게 변형되는 것을 알 수 있다. 이를 히스토그램으로 나타내면, 아래의 그림 5와 같다.

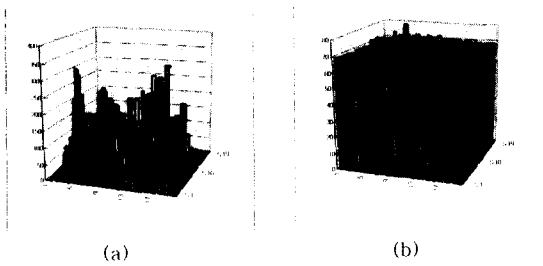


그림 5. 변형 후 분포

위의 그림에서와 같이 STR을 이용한 방법이 보다 균일한 분포를 얻을 수 있다.

4. 균등분포 변환과 고정 격자를 이용한 질의처리

본 장에서는 2장에서 설명한 고정격자방법을 균등분포로 전환된 공간객체에 적용하는 방법을 제안한다. 공간질의가 요청되면, 공간객체와 마찬가지로 공간질의의 위치도 변형되어야 한다. 공간질을 변형하는 방법은 탄성변형의 경우, [4]에 자세한 설명이 되어 있다. 또한 STR에 의한

변형도 유사한 방법으로 공간질의의 영역 위치를 변형할 수 있다. 이 두 가지 경우 모두, 질의영역의 변형을 위하여서는 변형을 위한 보조데이터가 필요하다. 그러나, 이 데이터는 크기가 작으므로 주기억장치에 있고, 질의영역의 변형을 위한 처리시간은 무시할 만큼 작다. 결과적으로, 변형된 공간질의의 영역으로 고정격자방식의 색인을 통하여 공간질의가 수행된다.

5. 결론

본 논문에서는 고정격자방법에 의한 공간색인 방법의 단순성과 효율성을 이용하기 위하여, 비균일적인 공간객체를 균일적인 공간객체의 분포로 전환하는 방법을 제안하였다. 또한 이 변환방법을 이용하여 기존의 고정격자방법의 공간색인을 효과적으로 이용하는 방법을 제안하였다.

이 방법은 다른 종류의 공간색인방법과 달리, 색인구조를 위한 디스크 접근이 필요 없고 단순히 데이터를 읽기 위한 디스크 접근 만이 필요하므로 성능이 매우 향상된다. 또한 색인방법이 단순하여, 구현 및 기존의 데이터베이스관리시스템에 통합하는 작업이 간편하여 진다.

그러나, 이 방법은 점 객체에 대하여서만 적용이 가능하다. 또한 아직까지는 2차원 공간의 객체에 대하여서만 적용을 하였다. 따라서, 이를 비점객체와 3차원 이상의 다차원 공간의 객체에 적용하는 연구가 계속하여 이루어져야 한다. 또한 본 논문에서 다루지 못한 성능에 관한 비교가 이루어져야 한다.

참고문헌

- [1] Beckmann N., Kriegel H. -P. and Schneider R., Seeger B., 'The R*-tree: An Efficient and Robust Access Method for Points and Rectangles', Proc. ACM SIGMOD, pp. 322-331, 1990
- [2] J. Nievergelt and H. Hinterberger, 'The Grid Files: An Adaptive, Symmetric Multikey File Structure', ACM TODS, vol. 9, No. 1, pp.38-71, 1984
- [3] Christos Faloutsos and Ibrahim Kamel, 'Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension.' PODS, pp.4-13, 1994
- [4] Jung-Rae Hwang, Ji-Hyun Oh, and Ki-Joune Li, 'A Query Transformation Method for Multi-Source Distributed Database Systems', Proc. 9th ACM-GIS Symposium, Atlanta, USA, 2001
- [5] Scott T. Leutenegger, J. M. Edgington and Mario A. Lopez, 'STR: A Simple and Efficient Algorithm for R-Tree Packing.', ICDE, pp.497-506, 1997