DTVF를 갖는 확장 R-tree 공간 색인 기법

정원일*, 정보흥*, 박동선*, 김재홍**, 배해영*
* 인하대학교 전자계산공학과, ** 영동공과대학교 컴퓨터공학과

Extended R-tree Spatial Indexing Methods with DTVF

Warn-Il Chung*, Bo-Heung Chung*, Dong-Seon Park*,
Jae-Hong Kim** and Hae-Young Bae*
* Dept. of Computer Science & Engineering, Inha University
** Dept. of Computer Engineering, Youngdong University

요 약

공간 인덱스를 이용한 공간 질의 처리의 과정은 여과와 정제 단계로 이뤄진다. 여과 단계에서 후보 객체의 수를 줄이면, 정제 단계에서의 false-hit이 낮아지므로 불필요한 디스크 접근과 공간 연산으로 인한 질의 처리 비용의 증대를 방지할 수 있다. 본 논문에서는 여과 단계에서 후보 객체를 최소화하기 위해 DTVF가 추가된 확장 R-tree를 제안한다. 제안된 기법에서는 n차원 상에 존재하는 공간 객체의 대표 성질들을 구석점 변환 기법을 이용하여 2n차원의 점으로 변환하고, 이 값을 확장된 R-tree 리프 노드의 DTVF에 유지한다. 공간 질의 처리시 여과 단계에서 DTVF를 이용하면 후보 객체 수를 최소화할 수 있으며, DTVF에 유지된 차원 변환된 값을 통해 후보 객체 선정에도 빠른 성능을 나타낸다. 제안된 기법은 공간 질의 처리시 여과 효율을 극대화하여 질의 처리 성능을 향상시킨다.

1. 서론

공간 데이터베이스 시스템의 등장과 이의 활용이 증대되면서 공간 데이터베이스를 위한 효과적인 질의 처리에 대한 연구가 활발히 진행되고 있다. 이러한 시스템에서의 질의 처리는 공간 질의 처리에 후보로 사용될 객체들을 선택하는 여과 단계와 이 단계를 통해 걸러진 객체들 중에서 실제 질의를 만족하는 객체들을 선정하는 정제 단계로 나누어 질의를 처리하는 방법이 연구되었다[1,2,3].

여과와 정제 방법을 이용한 공간 질의 처리를 위해 널리 사용되는 공간 색인은 R tree 계열이다. R-tree 계열의 공간 색인 기법은 해당 공간 객체의 최소 정계 사각형을 이용한다. 그러나 R-tree 계열에서 사용되는 최소 경계 사각형은 모든 공간 객체에 동일하게 적용됨으로써 둘리곤이나 폴리라인과 같은 공간 객체의 위래 형태를 왜이버리게 만든다. 이렇게 모든 객체에 획일적으로 적용된 최소 경계 사각형으로 인해 여과 단계에서 많은 후보 객체를 선정하게 되는 문제점을 안게 된다. 이러한 후보 객체는 실제 공간 객체의 기하학적 데이터를 다루는 정제 단계에서의 질의 처리 비용을 증대시켜 질의 처리 성능을 저하시키는 요인이 된다. 따라서 여과 단계에서 최소한의 후보 객체를 선정하여 정제 단계에서 야기되는 false-hit을 줄이는 것이 중요하다[1,3,4].

본 논문에서는 여과 단계에서 최소한의 후보 객체들을 선정하여 정제 단계에서의 질의 처리 비용 증대를 방지하기 위해 DTVF(Dimension Transformation Value Field)가 추가된 확장 R tree를 제안한다.

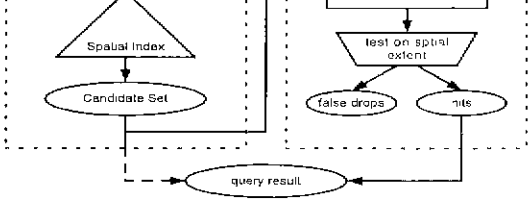
제안된 기법에서 R-tree의 리프 노드에 추가된 DTVF에는 공간 색인을 생성할 때 해당 공간 객체의 형태를 결정하는 주요한 정점들을 추출하고, 선택된 정점들을 구석점 변환 기법을 이용해 생성된 값이 유지된다. 구석점 변환 기법을 통해 얻어진 값으로부터 변환된 차원에서 공간 객체들간의 포함관계나 교차관계 등을 규정할 수 있다. 구성된 공간 색인을 통해 공간 질의를 처리할 때에는 해당 공간 객체의 최소 경계 사각형과 함께 DTVF의 값을이용하여 공간 객체간의 공간 관계를 확인하여 질의를 만족하는 객체와 후보 객체로 분류한다. 질의를 만족하는 객체는 질의 결과로 선택되고, 후보 객체는 정제 단계로 유도된다. 제안된 기법은 여과 단계에서 후보 객체를 산정할 때 정제 단계에서 false-hit을 야기하는 후보 객체를 사전에 배제함으로써 여과 효율을 극대화한다.

본 논문의 구성은 다음과 같다. 2장에서는 공간 질의 처리 방법, 공간 색인 기법 그리고 차원 변환 기법에 대해 살펴보고, 3장에서는 제안된 DTVF와 확장 R-tree에서의 색인 구조, 그리고 이를 이용한 공간 연산 처리 방법을 알아본다. 4장에서는 제안된 기법에 대한 실험 및 평가를 하고, 끝으로 5장에서 결론을 맺는다.

2. 관련연구

2.1 공간질의처리

공간 데이터베이스 시스템에서의 공간 질의 처리는 공간 질의를 만족하는 최소 정계 사각형을 갖는 모든 객체들을 찾아 후보 객체를 선정하는 여과 단계와 여과 단계에서 선택된 후보 객체들의 기하학적 데이터를 이용하여 실제 질의를 만족하는지를 확인하는 정제 단계를 통해 수행된다[1,2,3]. [그림2.1]은 공간 질의 처리의 수행 단계를 보여주고 있다[1].



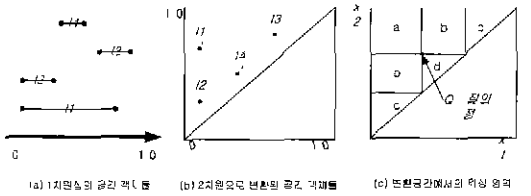
[그림2.1] 다단계 공간 질의 처리

2.2 R-tree

R-tree[1,3,4]는 B-tree를 n차원으로 확장한 색인 기법으로 공간 객체의 최소 경계 사각형을 이용하여 전체 변적이 최소화 되도록 구성된 트리이다. R-tree의 노드는 비리프 노드와 리프 노드로 이뤄진다. 리프 노드는 (I, tuple-identifier)로 표현되며, I는 해당 공간 객체가 차지하는 n차원 사각형을 나타내고 tuple-identifier는 데이터베이스의 튜플을 참조하는 인식자이다. 비리프 노드는 (I, child-pointer)로 표시되며, child-pointer는 R-tree의 하위 노드에 대한 주소를 나타내고 I는 하위 노드의 모든 엔트라들을 포함하는 사각형을 의미한다.

2.3 구석점 변환 기법

구석점 변환 기법[5,6]은 사각형 R의 좌하단 점을 (l_1, l_2), 우상단 점을 (u_1, u_2)라 할 때 이는 4차원 공간 상에서 한 점 (l_1, l_2, u_1, u_2)로 대응된다. [그림2.2]는 1차원상의 선분 *I1, I2, I3, I4*가 구석점 변환 기법에 의해 2차원상의 점 *I1', I2', I3', I4'*으로 변환됨을 보이고, 질의 점에 대한 공간 객체들간의 위상 영역을 나타내고 있다.



(a) 1차원상의 공간 객체 들 (b) 2차원으로 변환된 공간 객체들 (c) 변환공간에서의 위상 영역

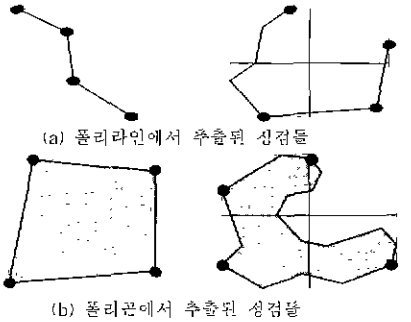
[그림 2.2] 구석점 변환 기법

변환 공간 상에 구축된 공간 객체에 대한 질의는 영역 교차 질의, 영역 포함 질의 그리고 영역 피포함 질의가 가능하다. 이러한 영역 질의를 처리하기 위해서는 주어진 질의 영역과 특성 공간 관계를 가지는 객체들이 변환 공간의 어느 부분에 위치하는 가를 알아야 한다. 질의 영역이 Q(q1, q2)로 주어졌을 경우 질의 영역과의 상대적 공간 관계에 따라 [그림2.2]와 같이 영역 a에서 영역 d까지 6개의 영역으로 구분할 수 있다. 영역 a는 원 공간에서 Q를 포함하는 모든 객체가 위치하고, 영역 b는 원 공간에서 Q에 교차하는 모든 객체가 위치하게 되며, 영역 c는 원 공간에서 Q와 전혀 겹치지 않는 모든 객체가 자리하며, 영역 d는 원 공간에서 Q에 포함되는 모든 객체가 위치하게 된다.

3. 확장 R-tree 색인 구조와 공간 연산 처리

3.1 DTVF

DTVF를 구성하기 위해서는 먼저 공간 객체를 이루고 있는 점들 중에서 구석점 변환 기법을 이용해 차원 변환할 점들을 즉 공간 객체를 대표할 수 있는 점점들을 선택한다. [그림3.1]의 (a)와 (b)에서 해당 공간 객체를 대표하는 점들을 보여주고 있다.



(a) 폴리라인에서 추출된 정점들

(b) 폴리곤에서 추출된 정점들

[그림3.1] 공간 객체를 결정하는 주요 정점 추출

위 그림에서의 공간 객체들에 대한 주요 정점을 추출하기 위한 규칙은 아래와 같다.

① 해당 공간 객체를 구성하는 정점이 4개 이하이면 모든 정점을 선택하고 정점 추출을 종료한다.

② 주어진 공간 객체의 영역을 동일한 크기로 4등분한다.

③ 한 영역에 점이 2개 이상이면 중심에서 가장 먼 점을 선택한다. 이때 모든 영역에 대해 점이 선택되었으면 정점 추출을 종료한다.

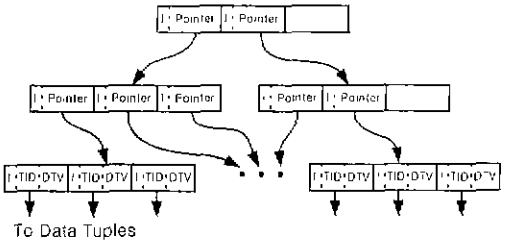
④ 공간 객체가 폴리곤이고 해당 영역에 점이 하나도 없으면, 대각 영역을 제외한 두 영역에서 각 선택된 점으로부터 가장 먼 거리에 있는 점을 갖이서 두 거리 중 먼 거리를 가지는 영역의 점을 선택하고 정점 추출을 종료한다.

⑤ 공간 객체가 폴리라인이고 해당 영역에 점이 존재하지 않으면, 선분의 겹침이 가장 큰 점들을 선택한다.

이렇게 추출된 공간 객체의 정점 4개를 구석점 변환시켜 생성된 값을 DVTF에 유지한다.

3.2 확장 R-tree 색인 구조

제안된 기법에서는 기존의 R-tree를 확장한 새로운 공간 색인을 이용한다. R-tree의 확장은 리프 노드만을 확장 변형하고 루트 노드나 중간 노드는 기존의 구조를 유지한다. 아래 [그림3.2]은 리프 노드를 확장한 R tree의 구조를 나타낸다.



[그림3.2] 확장 R-tree 구조

아래 [표1]에서 기존의 R-tree 노드와 확장된 R-tree의 노드의 변경사항을 확인할 수 있다.

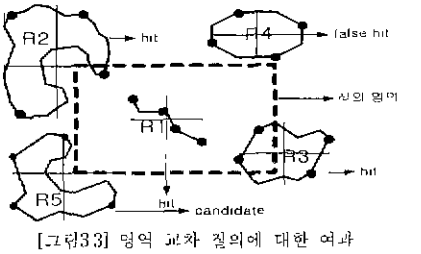
	R-tree	확장 R-tree
비리프 노드	(I, child-pointer)	(I, Pointer)
리프 노드	(I, tuple-identifier)	(I, TID, DTV)

[표1] R-tree와 확장 R-tree의 노드

확장 R-tree의 비리프 노드는 R-tree의 비리프 노드의 동일한 구조를 가지며, 확장 R-tree의 Pointer는 R-tree의 child-pointer이다. 또한, 확장 R-tree에서 리프 노드의 R-tree의 리프 노드에 DTV(Dimension Transformation Value) 필드를 추가한 형태를 취하고 있으며 여기서의 TID는 R-tree의 tuple-identifier이다. DTV는 공간 객체의 대표 정점들을 구석점 변환 기법을 적용하여 차원 변환시킨 값이다.

3.3 확장된 R-tree를 이용한 공간 연산 처리

제안된 색인 기법을 이용한 공간 질의 처리는 여과 단계에서 확장 R-tree에 유지되는 최소 경계 사각형과 DTVF를 동시에 적용한다. [그림3.2]는 질의 영역에 대해 영역 교차 질의를 처리하기 위한 여과 단계를 나타내고 있다. 공간 객체 R1은 질의 공간에 포함되므로 hit되고, R4는 질의 영역 외부에 존재하므로 false hit가 된다. R2, R3 그리고 R5는 MBR을 적용하면 후보 객체가 되지만 DTVF를 적용한 경우 R2와 R3는 질의 영역과 교차점을 얻을 수 있고, R5는 DTVF를 적용하고도 교차 여부를 알 수 없으므로 후보 객체로 선정된다.



[그림3.3] 영역 교차 질의에 대한 여과

제안된 확장 R-tree를 이용한 공간 질의 처리 알고리즘은 아래의 [알고리즘1]과 같다

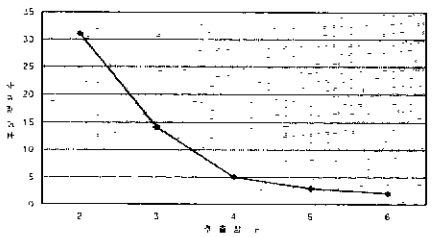
```
intersect(t, r)   ResultSet
INPUT
  t   extended R-tree
  r   query region
OUTPUT
  ResultSet,. all objects intersected with query region
{
  t' = t,
  r' = DimensionTransform(r'),
  ResultSet = FindObject(t', r, r'),
}
FindObject(t', r, r') {
  if node's MBR is intersected with that of r then
    if node is not leaf then FindObject(t', r, r'),
    else IsValidObject(t', r, r'),
  else false hit
}
IsValidObject(t', r, r') {
  if r is intersected with MBR of t' and r' is located in
  intersect area then hit
  else if r is intersected with MBR of t' and r' is not located
  in intersect area then candidate
  else if r' is beyond MBR of t' then false-hit
}
```

[알고리즘 1] 영역 교차 질의 처리

4. 성능 평가

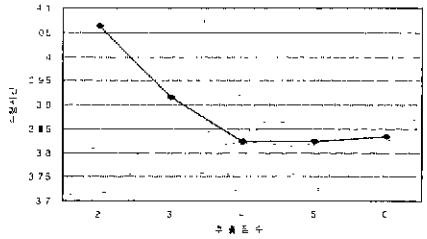
본 실험에서 사용된 질의는 질의 영역에 대해 교차되는 영역을 찾는 질의로 이는 [질의 예]에서 나타나 있다. 실험에서는 입력된 공간 질의에 대해 기존의 R-tree만을 이용한 여과와 확장 R-tree를 이용한 여과간의 여과율을 비교하였다. 실험에서 사용한 데이터는 질의의 대상이 되는 공간 데이터와 질의 영역에 대한 공간 데이터는 균일 분포, 정규 분포 등의 통계적인 방법을 사용하여 생성하였다. 모집단의 튜플 수는 100,000개, 질의 영역의 개수는 10,000개이며, 모집단 및 질의 영역의 위치 및 크기는 균일 분포 및 정규 분포에 따른 난수를 이용하였다.

[그림4.1]은 추출된 정점 수가 증가할 때 후보 객체 수의 변화에 대한 실험 결과로 추출되는 점의 수가 증가할수록 여과단계에서 선정되는 후보 객체의 수가 감소됨을 알 수 있다. 특히 추출점 수가 4보다 클 경우 급격하던 후보 객체수의 변화가 완만하게 나타남을 통해 추출 점의 개수를 4개 이상 선택해야함을 보여준다.



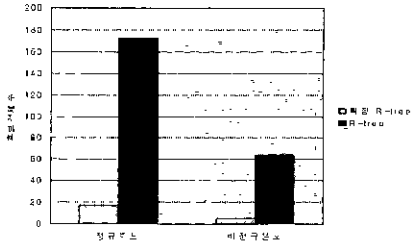
[그림4.1] 추출 정점 수와 후보 객체 수간의 비교

[그림4.2]는 추출점수와 수행시간과의 관계를 실험한 결과로 추출점의 수가 커짐에 따라 추출점수가 4일때까지는 감소를 보이고 이후는 오히려 수행시간이 증가한다. 이를 통해 추출점수가 4보다 작으면 후보 객체 수의 증가로 질의 처리 비용이 증대되고, 추출점수가 4보다 큰 경우에는 후보 객체 수는 감소했음에도 여과단계에서 후보객체를 선정하는 단계에서의 연산 비용 증가로 오히려 질의 처리 비용이 커짐을 알 수 있다. 따라서 공간 객체를 나타내는 점의 수는 4개가 이상적이다.



[그림4.2] 추출점수와 수행시간 비교

[그림4.3]은 모집단이 정규분포와 비정규분포를 따를 경우 산출되는 후보 객체의 수를 실험한 결과를 나타낸다. 이 실험으로부터 제안된 확장 R-tree를 이용하여 여과 단계를 수행할 경우 후보 객체 수가 현격하게 감소할 수 있음을 보여준다.



[그림4.3] 모집단 분포에 따른 후보 객체 산출

5. 결론

공간 데이터베이스 시스템에서의 공간 질의 처리는 여과와 정제 단계로 분류되고, 정제 단계에서 이루어지는 큰 질의 처리 비용을 감소시키기 위해서는 여과 단계에서 후보 객체의 수를 최소화해야 한다.

본 논문에서는 DTVF가 추가된 확장 R-tree를 이용하여 효율적인 공간 질의 처리를 수행하는 방법을 제안하였다. 제안된 기법은 공간 색인을 구성할 때 해당 공간 객체의 형상을 유지할 수 있는 주요 정점들을 선택하고, 추출된 이 점들을 구석점 변환 기법을 이용하여 차원 변환된 값을 DTVF에 유지하였다. 실제 질의 처리시 여과 단계에서 DTVF를 이용하여 후보 객체 수를 줄일 수 있었고, 이에 정제 단계에 입력되는 후보 객체 수가 최소화되므로 질의 처리 비용이 감소됨을 알 수 있다.

앞으로 연구해야 할 과제로는 민지 차원 변환을 위해 해당 공간 객체의 주요 정점들에 대한 추출 방법에 대한 연구와 구석점 변환 기법을 통해 차원 변환된 값들의 효율적인 관리 방법에 대한 연구가 필요하다.

참고 문헌

[1] Volker Gaede and Oliver Gunther, "Multidimensional Access Methods" ACM Computing Surveys, Vol 30, No 2, June 1998

[2] T. Brinkhoff, H P Kriegel and R Schneider, Efficient Spatial Query Processing in Geographic Database Systems, IEEE Data Engineering Bulletin, Vol 16, No 3, pp 10-15, 1993

[3] T. Brinkhoff, H P Kriegel and B Seeger, "Efficient Processing of Spatial joins using R-trees," in proc. Intl Conf on Management of Data, ACM SIGMOD, pp 237-246, May 1993

[4] Antonin Guttman, "R-trees : A Dynamic Index Structure For Spatial Searching," Proceedings of ACM SIGMOD International Conference on Management of Data, Boston, MA, pp 47-57, 1984

[5] B. Seeger and H. P. Kriegal, "Techniques for Design and Implementation of Efficient Spatial Access Methods," In Proc 14th Intl Conf on Very Large Data Bases, pp 360-371 1988

[6] K. Hinrich and J Nievergelt "The Grid File : A Data Structure Designed to Support Proximity Queries on Spatial Objects," In Proc Intl Workshop on Graph Theoretic Concepts in Computer Science, pp 100-113, 1983

[7] B U Pagel, H W Six, and H Toben, "The Transformation Technique for Spatial Objects Revisited," In Proc 3rd Intl Symp on Spatial Databases, SSD93, 1993