튜 토 리 얼 Ⅲ

공간 데이터베이스 연구 동향

-공간 색인과 공간 질의 처리를 중심으로-

▷ 연 사 : **송 주 원** 박사 (한국통신 멀티미디어 연구소)

▷ 사 회 : **이 경 현** 교수 (부경대)

공간 데이터베이스 연구 동향
"구석점 변환 기법을 이용한 공간 질의 처리"를 중심으로

2000. 5.

한국통신 멀티미디어연구소
송주원

목차

- 서론
  - 공간 데이터베이스 시스템
  - 공간 액세스 방법
- 주요 공헌
- 구석점 변환 기법
  - 변환 객체의 분포 특성
  - 변환공간의 영역 구분
- 공간 조인 알고리즘
  - 정의 및 예
  - 기존 연구들
  - 공간 조인 윈도우(SJW)
  - SJW 간의 관계
  - 알고리즘의 기본 골격
  - 최적화 순서
  - 이차원 확장
  - 공간 조인 비교 실험
- 클러스터링 특성
  - 변환공간의 객체 분포 특징
  - 클러스터링 기준
  - 클러스터링 비교 실험
- 결론
- 참고문헌

서론

- 공간 데이타베이스 시스템
  - 공간 객체를 관리하는 데이타베이스 시스템
    - 공간 객체는 공간 데이터와 비공간 데이터를 가짐
    - 이들 데이터를 유기적으로 관리하는 구조를 가져야 함
  - 지리 정보 시스템 등의 기본 프레임워크로 활용됨
  - 공간 질의를 효율적으로 처리하기 위하여 저장 시스템 내에 효율적인(클러스터링을 유지하는) 공간 액세스 방법이 색인으로 제공되어야 함
  - 또한, 저장 시스템은 공간 객체의 두 종류 데이터인 공간 데이터와 비공간 데이터가 유기적으로 관리할 수 있는 구조를 가져야 함
    - 공간 데이터: 객체의 위치, 크기, 모양, 다른 공간 객체와의 공간적인 상호 관계 등의 데이터
    - 비공간 데이터: 문자, 숫자 등으로 표현될 수 있는 일반 데이터
  - 효율적인 공간 조인 및 공간 질의 처리 알고리즘이 제공되어야 함

서론

- 공간 데이터베이스(지리 정보 시스템)에서는 여러 형태의 질의가 효율적으로 처리될 수 있어야 함
  - 공간 질의(spatial query)
  - 비공간 질의(non-spatial query)
  - 복합 질의
- 공간 질의
  - 기본 형태
    - 영역질의: 영역 포함 질의, 영역 피포함 질의, 영역 교차 질의, 점질의
    - 최근접 이웃(nearest neighbor) 질의
    - 연결 선분(connected line segments) 질의
  - 공간 조인

서론

- 공간 데이터베이스 분야에서의 중요 연구 이슈
  - 공간 액세스 방법(Spatial Access Method: SAM)
    - clustering, performance, concurrency, recovery
  - SAM을 이용하는 공간 질의 및 조인 처리 알고리즘
    - 필터링 및 정련화의 이단계 처리 방법 이용
  - 저장 시스템 확장: SAM을 포함
  - 공간 데이터 모델링: 객체 지향 개념 이용 등
  - 공간 질의어: SQL 확장 등
  - 사용자 인터페이스

** 크기를 가지는 공간 객체를 다루는 데 있어서 일반 객체를 다루는 방법을 어떻게 확장하면 될 까

** 객체의 크기 때문에 특별히 생기는 문제는 무엇일 까

GIS 아키텍춰 발전 단계

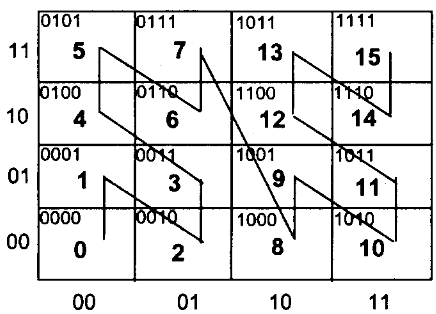
- 단계별 기반 시스템
  - 제 1단계: 화일 시스템
  - 제 2단계: 이원형(Dual) 아키텍춰
    - 공간 데이터는 화일 시스템, 비공간 데이터는 관계형 DBMS에 저장
  - 제 3단계: 계층형(Layered) 아키텍춰
    - 두 데이터 모두 관계형 DBMS에 저장
    - 공간 데이터를 위한 기능은 RDBMS 위의 공간 기능 확장기 계층에서 제공
  - 제 4단계: 공간 데이터베이스 시스템
  - 제 5단계: 공간 및 객체 지향 데이터베이스 시스템의 통합 시스템 --> 객체 지향 GIS 엔진 ex) GEOSS, GEUS

서론

- 공간 액세스 방법(spatial access method: SAM)
  - 공간 객체들을 관리하는 액세스 방법으로 공간 데이타베이스 시스템에서 (클러스터링) 색인으로 사용됨
  - 기존의 다차원 점 액세스 방법(Point Access Method: PAM)에 객체의 크기를 관리할 수 있는 방법 추가
  - 객체의 색인을 위한 키로 포함자(container) 이용
    - 대표적 포함자: MBR(Minimum Bounding Rectangle)
  - 네 부류: 공간 순서화 기법, 객체 분할 또는 중복 기법, 영역 겹침 기법, 변환 기법
  - SAM은 객체의 공간 데이타에 대한 클러스터링 성질을 가져야 함
    - 인접성 기반

서론

- 공간 순서화 기법
  - 다차원 공간을 서로 소(disjoint)인 영역들로 나누고 space filling curve(예, z-order)를 이용하여 각 영역을 일차원으로 순서화한 후 객체와 교차하는 영역들에 그 객체에 대한 정보를 현존 RDBMS의 일차원 액세스 방법(예, B+-tree)을 이용하여 저장함
  - z-order [Peano1908; Morton1966; Orenstein1983]



서론

- 객체 분할 또는 중복 기법
  - 전체 도메인 영역을 서로 소인 영역들로 나누고 공간 객체와 교차되는 모든 영역에 객체 정보를 저장(기존의 다차원 점 액세스 방법을 그대로 이용)
    - clipping: 객체의 해당 부분 정보를 교차되는 영역별로 각각 저장함, 부분 정보의 통합 방안이 요구됨
    - duplication: 객체 전체의 정보를 교차되는 모든 영역에 중복 저장
    - 예: R+-tree[Sellis87], Cell-tree[Günther91]
- 영역 겹침 기법
  - 공간 객체에 대한 포함자 간에 영역 겹침을 허용
  - 효율성을 위하여 이 겹침을 최소화하는 방법들이 연구됨
  - 예: R-tree[Guttman84], R*-tree[Beckman90], skd-tree[Ooi89]

서론

- 변환 기법
  - 원공간에서 크기를 가지는 객체를 원공간보다 높은 차원을 가지는 변환공간의 점으로 표현함
  - 변환된 객체는 다차원의 점 액세스 방법(point access method: PAM)으로 관리함
  - 대표적인 변환 기법에는 구석점 변환 기법과 중앙점 변환 기법이 있음
  - 80년대에는 클러스터링 성질을 유지하지 못한다고 주장되어 왔음

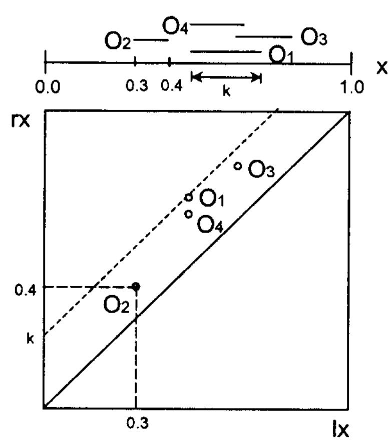
연구의 주요 공헌

- 구석점 변환 기법은 클러스터링 성질을 유지함을 보임
- 구석점 변환 기법 기반 SAM의 하나로 MBR-MLGF를 제안함
- 구석점 변환 기법을 이용하는 세계 최초의 효율적인 공간 질의 처리 알고리즘을 제안함
- 구석점 변환 기법을 공간 데이타베이스 시스템에서 클러스터링 색인으로 사용할 수 있는 방법을 제안함
- GEOSS 개발을 위한 이론적 근거를 제시함

구석점 변환기법

- n차원 원공간의 객체에 대한 최소 포함 사각형 (minimum bounding rectangle: MBR)의 각 차원에 대한 최소값, 최대값을 이용하여 2n 차원의 점으로 표현
- 일차원 원공간의 경우
  - 일차원 원공간의 선분 (<lx>, <rx>) ---> 이차원 변환공간의 점 (<lx, rx>)

- Example:

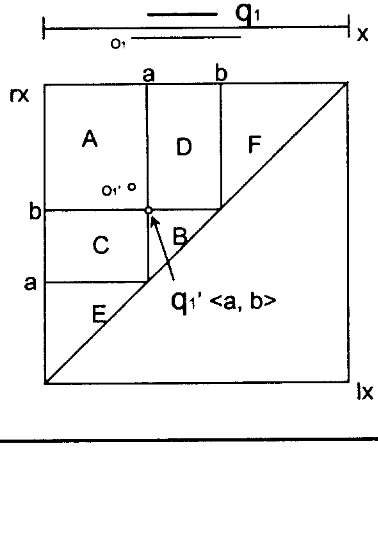


변환 객체의 분포 특성

- 객체는 대각선 위쪽 부분에만 분포
- 객체의 최대 크기(k)가 정해져 있을 경우 객체는 대각선 위의 폭이 좁은 띠 내부에 분포 -> highly skewed distribution(불균일 분포)
- 이러한 불균일 분포를 효율적으로 관리할 수 있는 PAM이 필요
  - 예: MLGF(Whang and Krishnamurthy), LSD-tree (Henrich), Buddy-tree(Seeger)

변환공간의 영역 구분

- 주어진 질의 영역 q_1에 대한 공간 관계에 따라 변환공간은 A에서 F까지의 여섯개의 영역으로 구분됨



A: 원공간에서 q_1을 포함하는 객체들이 존재하는 영역
B: 원공간에서 q_1에 포함되는 객체들이 존재하는 영역
C: 원공간에서 q_1의 왼쪽 점과 교차하는 객체들이 존재하는 영역
D: 원공간에서 q_1의 오른쪽 점과 교차하는 객체들이 존재하는 영역
E: 원공간에서 q_1의 왼쪽 편에 위치하는 객체들이 존재하는 영역
F: 원공간에서 q_1의 오른쪽 편에 위치하는 객체들이 존재하는 영역

- q_1에 대한 영역 교차 질의 처리: A U B U C U D

공간 조인 알고리즘

- 공간 조인 $R \bowtie_{i\theta j} S$
  - 공간 관계 θ 를 가지는 공간 객체 쌍들을 찾는 연산

R, S: 공간 객체 클래스(또는 화일)

θ: 공간 관계 연산자

- 예: Within 10 km from, To the Northewest of

i: R의 i번째 (공간) 속성

j: S의 j번째 (공간) 속성

- 예: points, lines, polygons

공간 조인의 예

- house(identification, price, location)
- lake(identification, name, area)
  - location: type point
  - area: type polygon

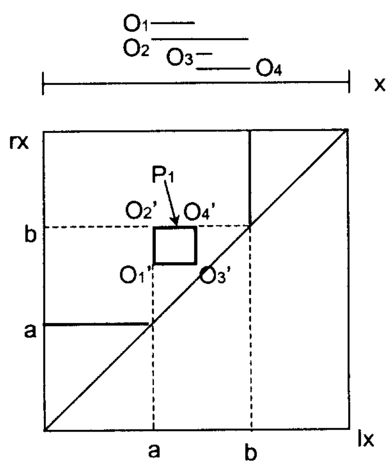
- 질의: Find all houses within 10 km from a lake
  - house $\bowtie_{3\theta 3}$ lake
  - where θ : within 10 km from

기존 연구들

- R-tree 기반 알고리즘
  - Using the R*-tree [Brinkoff et. al., '93]
  - Using the R*-tree with Breath-first search[Huang & Jing '97]
  - Using the generalization tree [Günther '93]
    - an abstraction of the R-tree
  - Using the seeded tree [Lo & Ravishankar '94]
    - for temporary files
- not SAM-based
  - Partition-based Spatial-Merge Join[Patel '96]
  - Spatial Hash Join[Lo & Ravishankar '96]
- Linearization-order (space filling curve) 기반 알고리즘
  - Using the z-order and the B-tree [Orenstein '86]
- 변환 기법 기반 알고리즘
  - None has been proposed

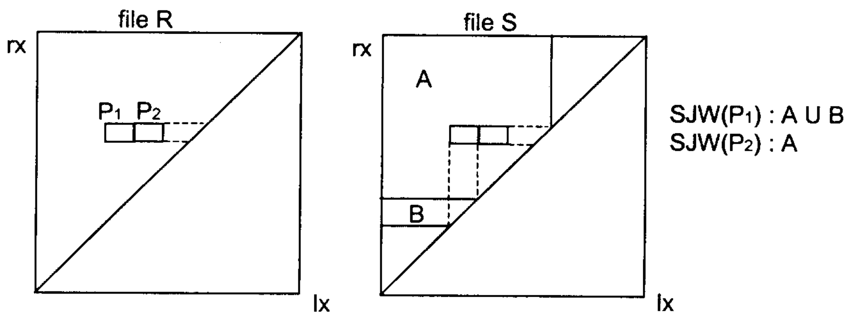
공간 조인 윈도우

- 영역 P_1 에 대한 공간 조인 윈도우(spatial join window: SJW): P_1과 교차하는 객체들이 존재할 수 있는 최소한의 영역
- 영역 P_1의 좌상점(O_2')의 좌표값이 P_1의 SJW를 결정함



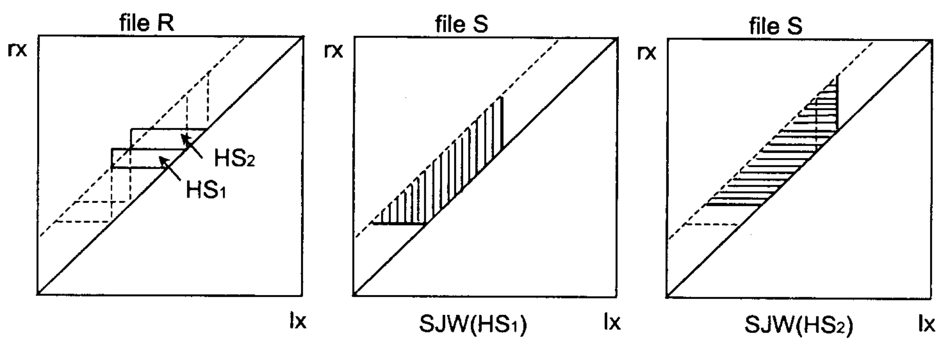
SJW 간의 관계(I)

- 같은 띠 내의 두 영역에 대한 SJW 간의 관계:
  - 대각선에서 먼 영역에 대한 SJW는 가까운 영역에 대한 SJW 를 포함함



SJW 간의 관계(II)

- 인접한 두 띠에 대한 SJW 간의 관계:
  - 아래와 오른 편 일부를 제외하고는 대부분이 겹침



공간 조인 알고리즘의 기본 골격

- 기본 골격

for each horizontal strip *i_strip*

for each grid region *e_region* in SJW of *i_strip*

for each grid region *i_region* of *i_strip*

if *e_region* is in SJW of *i_region*

do join(*e_region*, *i_region*)

- 알고리즘의 최적화
  - 가정: LRU buffer replacement policy
  - 정책: 버퍼에 남아있는 페이지를 먼저 처리(디스크 I/O 최소화)

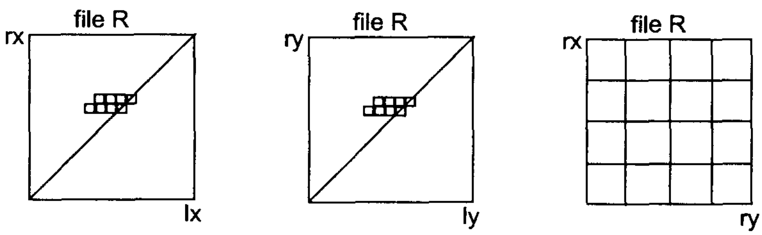
최적화 순서

- 띠들의 처리 순서 (Outer loop)
  - 띠들에 대한 SJW 간의 차이를 줄이기 위하여 인접한 띠를 선택
  - bottom-up order 또는 top-down order
- SJW 내의 그리드 영역들의 처리 순서 (Middle loop)
  - 한 띠에 대한 SJW의 그리드 영역들: row-major, bottom-up
  - 다음 띠에 대한 SJW의 그리드 영역들: row-major, top-down
- 띠 내부의 그리드 영역들의 처리 순서 (Inner loop)
  - SJW의 한 그리드 영역에 대한 띠내의 그리드 영역들: far-near order
  - SJW의 다음 그리드 영역에 대한 띠내의 그리드 영역들 : near-far order

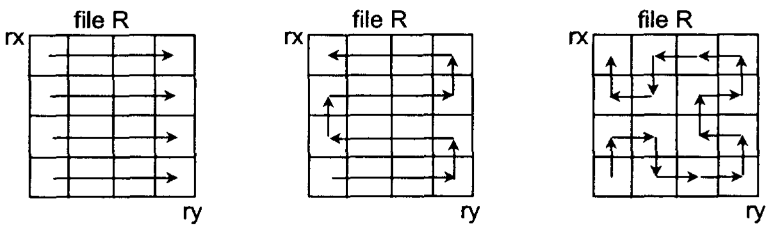
이차원 알고리즘

- 알고리즘의 골격
  - 일차원의 경우와 유사
- Issues
  - 사차원 변환 공간에서의 초월띠의 선택(hyper-strip)
  - 초월띠의 처리 순서
    - SJW 간의 차이를 줄이는 순서
  - SJW 내의 그리드 영역 처리 순서
  - 띠 내의 그리드 영역 처리 순서

초월띠의 선택(I)



- Each region in ry-rx plane is a two-dim. strip



row-major order　row-prime order　Peano-Hilbert order

초월띠의 선택(II)

- 우리는 Peano-Hilbert order를 이용
- 그 이유
  - SJW 간의 변화를 최소화하려면 가상띠들은 서로 인접하여야 함
  - Peano-Hilbert order는 재귀적이므로 MLGF와 같은 계층적 구조에 쉽게 적용될 수 있음

공간 조인 비교 실험

- 목적
  - R*-tree 기반 알고리즘과의 비교 실험을 통하여 제안된 알고리즘의 우수성을 입증하고자 함
- 실험 모델
  - 실험 데이타
    - 실제 데이타: 캘리포니아 지역의 도로망(131,461개) 및 강과 철도(128,971개) 데이타 집합
      - 86,094쌍이 교차(교차율: $5.077897*10^{-6}$)
    - 생성 데이타
      - 중앙점 균일 분포(U1, U2) 및 지수 분포(E1, E2)
      - 각 축에서의 길이: 최대 길이(k) 범위 내에서 균일 분포
      - 갯수: 각각 13만개
      - 교차율: $5.1736094*10^{-6}$ 및 $5.1684024*10^{-6}$
- MBR-MLGF와 R*-tree의 최대 블로킹 팩터는 동일

실험 결과(정규화)

실제 데이타

제목:

작성한 사람:
gnuplot
미리 보기:
이 EPS 그림은 미리 보기 그림을 포함하지 않고
저장되었습니다.
설명:
이 EPS 그림은 PostScript 프린터를
제외한 다른 프린터에서는
인쇄되지 않습니다.

생성 데이타

제목:

작성한 사람:
gnuplot
미리 보기:
이 EPS 그림은 미리 보기 그림을 포함하지 않고
저장되었습니다.
설명:
이 EPS 그림은 PostScript 프린터를
제외한 다른 프린터에서는
인쇄되지 않습니다.

실험 결과

- 공간 조인 비교 실험 결과 실제 데이타와 제작 데이타에 대한 MBR-MLGF를 이용한 제안된 알고리즘의 성능은 R*-tree 기반 알고리즘보다 우수함
- 이유
  - 제안 알고리즘: 전체적인(global) 최적화 방법 이용
    - SJW 간의 관계, LRU 버퍼링 방법의 특성을 고려한 때, SJW 내의 페이지 액세스 순서 결정
  - R*-tree 기반 알고리즘: 공간적 국부성(spatial locality) 만을 이용
    - 조인되는 두 노드 내의 디렉토리 엔트리들이가리키는 페이지들에 대한 액세스 순서 만을 고려함
    - local z-ordering, plane sweeping, plane-sweeping with pinning 등의 방법 이용

변환 기법의 클러스터링 특성

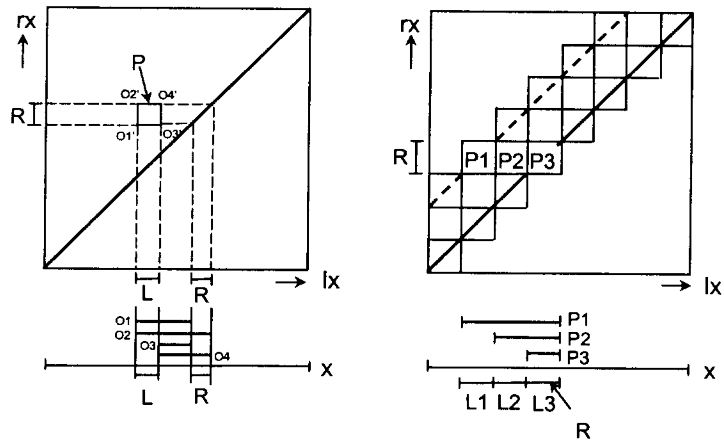
- 클러스터링: 객체의 어떤 속성(공간 데이타)이 유사한 객체들을 공간 상의 가까운 위치에 모아주는 성질
  - 클러스터링 성질이 유지되면 이 속성에 대한 질의를 효율적으로 처리할 수 있음
- 변환 기법은 클러스터링 성질을 유지하지 못한다고 주장되어 왔음
  - 근거: 원공간에서 인접한 객체들이 변환공간에서 인접하지 않은 위치에 존재함
  - 클러스터링의 기준: 객체 간의 최소 거리(인접성)
  - 문제점: 크기가 있는 객체들을 원공간에서 점으로 표현되는 객체의 경우와 마찬가지로 거리의 인접성 만을 클러스터링 기준으로 고려해야 하는가?

객체간 최소거리 클러스터링 기준

- 문제점
  - 객체들이 일부분 교차할 경우: 교차되는 부분의 크기에 관계없이 객체들 간의 키 값의 유사성은 같은가?
  - 객체가 교차하지 않는 경우: 객체간 최소거리가 동일한 경우 객체의 크기에 관계 없이 키 값의 유사성은 같은가?
- 객체의 크기는 고려하지 않고 최소 거리 만을 기준으로 고려하기 때문에 생기는 문제임

변환공간의 객체 분포 특징

- 구석점 변환 기법은 MBR의 크기와 위치가 유사한 객체들을 변환공간의 인접한 곳에 위치시킴



구석점 변환기법의 클러스터링 기준

- 구석점 변환 기법은 객체의 위치와 크기를 동시에 고려하는 클러스터링 기준을 사용함
- 이 기준이 의미있는 이유
  - 공간 액세스 방법에서 MBR은 키값으로 사용됨
  - MBR의 크기와 위치가 유사한 객체들(키값이 유사한 객체들)은 같은 질의에 의하여 동시에 액세스될 가능성이 높음
  - 객체 간의 최소 거리(인접성) 기준보다 의미있는 기준임
- 구석점 변환 기법은 클러스터링 성질을 유지하는 특성을 가짐

클러스터링 비교 실험

- 목적
  - MBR-MLGF와 R*-tree의 비교 실험을 통하여 MBR-MLGF가 클러스터링 성질을 유지함을 보임
- MBR-MLGF의 특징
  - MLGF의 구조를 변환공간의 각 분할 영역에 존재하는 객체들에 대한 원공간에서의 MBR을 각 영역에 대응하는 엔트리에 추가적으로 유지하도록 개선함
  - MBR 추가 유지로 필요한 저장공간 오버헤드는 없음
  - MLGF보다 영역 질의 처리에 대한 성능이 향상됨

실험 모델

- 실험 데이타
  - 실제 데이타: 캘리포니아 도로 객체에 대한 MBR(53,145개)
  - 생성 데이타
    - MBR 갯수: 50,000개
    - 중앙점 분포: 균일 분포, 지수 분포
    - 데이타 밀집도(d): 전체영역에 대한 MBR들의 면적 합의 비율, 0.5, 1.0, .., 3.0(여섯 경우)
    - 각축에 대한 길이 분포: 최대 길이(k) 범위 내에서 균일 분포
    - 최대 길이는 데이타 밀집도에 따라 계산
    - 총 12개 집합을 대상
- 질의 데이타
  - 질의 종류: 영역 교차 질의, 영역 포함 질의
  - 질의 영역: 전체 영역에 대한 면적 비율이 0.01, 0.1, 1, 2, 3, ...,10(%)(12 경우)인 정사각형(각 500개)
- MBR-MLGF와 R*-tree의 블록킹 팩터는 동일

실험 결과(정규화)

균일 분포(d=0.5)

제목:

작성한 사람:
gnuplot
미리 보기:
이 EPS 그림은 미리 보기 그림을 포함하지 않고
저장되었습니다.
설명:
이 EPS 그림은 PostScript 프린터를
제외한 다른 프린터에서는
인쇄되지 않습니다.

지수 분포(d=1.0)

제목:

작성한 사람:
gnuplot
미리 보기:
이 EPS 그림은 미리 보기 그림을 포함하지 않고
저장되었습니다.
설명:
이 EPS 그림은 PostScript 프린터를
제외한 다른 프린터에서는
인쇄되지 않습니다.

실험 결과 (정규화)

실제 데이타

제목:

작성한 사람:
gnuplot
미리 보기:
이 EPS 그림은 미리 보기 그림을 포함하지 않고
저장되었습니다.
설명:
이 EPS 그림은 PostScript 프린터를
제외한 다른 프린터에서는
인쇄되지 않습니다.

- 결과 요약
  - 실제 데이타와 제작 데이타에 대한 정규화된 영역 질의 실험 결과 MBR-MLGF와 R*-tree의 성능은 유사함
  - R*-tree는 클러스터링 성질을 유지한다고 알려져 있으므로 이 결과는 MBR-MLGF가 클러스터링 성질을 유지함을 보이는 것임

결론(I)

- 연구 결과 요약
  - 구석점 변환 기법은 클러스터링 성질을 유지함을 보임
  - 구석점 변환 기법 기반 SAM인 MBR-MLGF를 제안함
    - MBR-MLGF가 클러스터링 성질을 유지함을 실험을 통하여 보임
  - 공간 조인을 포함한 여러 공간 질의 처리를 위한 효율적인 알고리즘들을 제안함
    - 제안된 공간 조인 알고리즘과 R*-tree 기반 공간 조인 알고리즘과의 비교 실험을 통하여 제안 알고리즘의 우수성을 입증함
  - 구석점 변환 기법을 공간 데이타베이스 시스템에서 클러스터링 색인으로 사용할 수 있는 방법을 제안함

결론(II)

- 구석점 변환 기법은 공간 데이타베이스 시스템에서 실제적으로 색인으로 사용될 수 있는 SAM의 한 부류임
  - KAOSS에 MBR-MLGF를 공간 색인으로 사용하도록 확장한 공간 객체 저장 시스템 GEOSS에 이러한 연구 결과를 적용하였음
- 본 연구의 의의
  - "변환 기법을 이용한 공간 질의 처리"라는 새로운 연구 분야를 여는 중요한 연구임

참고문헌

- 황규영, 송주원, "대형 지리 정보 데이타베이스를 위한 객체지향 GIS 엔진," *정보과학회지*, 제13권 제3호, pp.77-87, 1995년 3월.
- 송주원, 김상욱, 황규영, "구석점 변환 기법을 이용한 공간 조인 알고리즘," *한국정보과학회 논문지*, 제 23권 제7호, pp. 682-698, 1996년 7월.
- 송주원, 이영구, 김상욱, 황규영, "공간 데이타베이스에서 구석점 변환 기법의 클러스터링 성질," *한국정보과학회 논문지*(B), 제24권 8호, pp. 797-809, 1997년 8월.
- Ju-Won Song et.al., "Spatial Join Processing Using Corner Transformation," *IEEE Trans. on Knowledge and Data Engineering*, Vol11. No.4, pp. 688-695, July/Aug. 1999.
- Ju-Won Song, et.al., "Transformation-Based Spatial Join," In *Proc. 8th Intl. Conf. On Information Knowledge Management* (*CIKM '99*), pp.15-26, Nov. 1999.
- Ju-Won Song, et al, "The Clustering Property of Corner Transformation in Spatial Database Applications", In Proc. 23th Intl. Computer Software and Applications Conf. (*COMPSAC '99*), Oct. 1999.