정보처리학회 논문지 제5권 제1호(1998.1)

최적탐색거리를 이용한 최소근접질의 처리 방법의 성능 평가

선 휘 준† · 김 홍 기††

요 약

공간 데이타베이스 시스템에서 최소근접질의는 매우 빈번히 발생하며, 다른 유형의 공간질의에 비하여 처리비용이 많이 요구된다. 최소근접질의의 처리비용을 최적화하기 위해서는 색인에서 검색되는 노드의 수를 최소화할 수 있어야 한다. 이를 위해 최소근접질의 처리시 색인에서 방문될 노드들을 정확히 선정하기 위한 검색거리 측도인 최적탐색거리가 제안되었다[13]. 본 논문에서는 최적탐색거리의 특성을 N차원으로 확장하고 최대검색거리를 이용한 방법에 비해 최적탐색거리를 이용한 방법이 질의처리 성능이 더 우수함을 실험을 통하여 입증한다.

The Performance Evaluation of Method to Process Nearest Neighbor Queries Using an Optimal Search Distance

Hwi-Joon Seon† · Hong-Ki Kim††

ABSTRACT

In spatial database systems, the nearest neighbor query occurs frequently and requires the processing cost higher than other spatial queries do. The number of nodes to be searched in the index can be minimized for optimizing the cost of processing the nearest neighbor query. The optimal search distance is proposed for the measurement of a search distance to accurately select the nodes which will be searched in the nearest neighbor query. In this paper, we prove properties of the optimal search distance in N-dimensional. We show through experiments that the performance of query processing of our method is superior to other method using maximum search distance.

1. 서 론

공간 데이타베이스 시스템은 컴퓨터지원설계, 지리정보시스템, 화상처리, VLSI 설계 등과 같은 응용분야에서 취급되는 공간적인 성질을 갖는 데이타를 효율적으로 관리하기 위한 시스템이다[6, 9]. 이러한 시스템의 전체적인 성능을 향상시키기 위해서는 공간질의의 효율적인 처리방법이 필요하다[2, 3, 4]. 공간 데이타베이스 시스템에서 취급되는 여러 유형의 공간질의들 중 주어진 위치에서 가장 가까운 공간객체(이하 객체라 함)를 찾는 최소근접질의(nearest neighbor query)는 매우 빈번히 발생하며, 최소근접질의의 처리는 연산 및 보조기억장치 접근을 위한 많은 처리 시간이 요구된다.

본 논문에서는 객체를 최소경계사각형으로 표현하는 공간색인방법들[1, 5] 중에서 R-트리를 이용한 최소근접질의 처리방법을 고려한다. R-트리에서 최소근접질

† 정 회 원 : 서남대학교 전산정보학과 교수
†† 정 회 원 : 동신대학교 전산통계학과 교수
논문접수 : 1998년 8월 5일, 심사완료 : 1998년 11월 19일

의를 처리하기 위한 기존의 방법들은 질의 기준이 되는 객체 또는 위치가 점일 경우만 적용 가능하거나, 2차원 검색공간에서 발생하는 최소근접질의 처리만을 고려하였다. 또한 객체 또는 부검색공간이 차지하는 영역을 나타내는 최소경계사각형들 간의 겹침을 고려하지 않았으며, 색인에서 검색되는 노드의 수를 정확히 줄이지는 못했다[13]. [13]에서는 최소근접질의를 처리할 경우 색인에서 방문되는 노드의 수를 최소로 하기 위한 검색거리 측도인 최적탐색거리가 제안되었으며, 이에 따른 특성 및 정확성이 2차원 검색공간에서 정리되었다.

본 논문에서는 최적탐색거리의 특성을 N차원 검색공간으로 확장하여 그 특성을 정리하였다. 그리고 실험에서는 최적탐색거리를 R-트리에 적용하여 기존의 방법과 최소근접질의에 따른 처리 비용을 비교 평가하였다. 질의 처리 비용을 비교 평가하기 위한 지수로는 질의처리시간, 디스크 접근 횟수를 이용하였으며, 객체들이 검색공간에 균일하게 발생되는 경우와 일부분에 집중되어 발생되는 경우 최소근접질의의 처리 성능을 규명하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로서 기존의 최소근접질의 처리방법에서 발생되는 문제점들을 간략히 알아본다. 3장에서는 논의의 대상이 되는 확장최소거리와 최적탐색거리에 대해 기술한다. 그리고 최소근접질의 처리시 최적탐색거리가 기존에 제시된 최대검색거리[14]보다 질의처리의 결과가 더 정확함을 N차원 검색공간으로 확장하여 증명한다. 4장에서는 실험을 통하여 최적탐색거리를 적용한 최소근접질의의 처리 성능을 기존의 방법과 비교 분석하고, 끝으로 5장에서는 결론을 내린다.

2. 관련연구

Roussopoulos는 R-트리를 이용하여 주어진 질의 기준이 되는 객체 또는 위치가 점일 경우 가장 가까운 다각형 객체를 찾는 방법을 제시하였다[12]. 이 방법은 다차원 검색공간에서 주어진 질의 점으로부터 객체 또는 부검색공간을 나타내는 최소경계사각형까지의 최소거리(minimum distance : MINDIST)와 최대검색거리(minimax distance : MINMAXDIST)를 계산하여 트리에서 검색되는 노드의 수를 줄이고자 하였다.

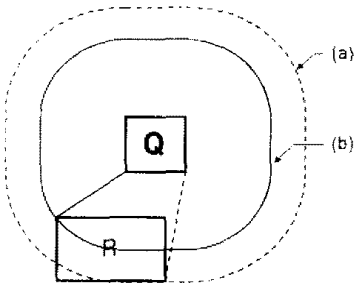
MINDIST는 질의 점 P에서 최소경계사각형 R의 각 경계 사각형 변 또는 꼭지점까지의 최소거리이다. 그리고 MINMAXDIST는 질의 점 P와 객체를 포함하고 있는 최소경계사각형 R과의 가능한 최대거리들 중에서 최소거리이다. 즉, P로부터 R을 구성하는 모든 축에서 가장 먼 점까지의 거리들 중에서 최소가 되는 거리를 의미한다.

제시된 방법에서는 최소근접질의 처리시 질의 기준이 되는 객체 또는 위치가 점일 경우만을 고려하였으며, 동적인 환경에서 발생하는 최소근접질의에 따른 성능평가가 이루어지지 않았다[13].

[14]에서는 질의 기준이 사각형일 경우 R-트리를 이용한 최소근접질의 처리방법을 제시하였다. 이 방법은 2차원 검색공간에서 최소근접질의 처리시 검색되는 노드들의 수를 줄이기 위해, [12]에서 제시된 최소거리(MINDIST)와 최대검색거리(MINMAXDIST) 개념을 확장하여 이용하였다. 이 방법은 두 최소경계사각형 사이에 가능한 객체간의 최대거리를 이용하여 R-트리에서 검색되는 노드의 수를 줄이고자 하였으나, 다음과 같은 문제점이 있다.

첫째, 2차원 검색공간에서 발생하는 최소근접질의 처리만을 전제로 하였다. 그러나 공간 데이타베이스 시스템에서는 다차원 검색공간에서 발생하는 최소근접질의 처리 방법이 요구된다.

둘째, 두 개의 최소경계사각형에 따른 부검색공간에 포함된 객체 또는 부검색공간들 간의 가능한 최대거리를 구하기 위해 최소경계사각형 간에 마주하는 대각선상의 꼭지점을 이용하기 때문에 오차를 포함하고 있다. 따라서 트리의 하위레벨로 내려갈수록 불필요한 노드들의 검색이 더욱 증가되므로 이 방법은 최소근접질의 처리시 검색되는 노드들의 수를 정확히 줄이지 못한 방법이다. 그림 1은 2차원 검색공간에서 질의 기준 Q로부터 최소경계사각형 R에 반드시 객체 및 부검색공간이 존재할 수 있는 거리를 나타낸 것이다. 그림 1에서 (a)는 MINMAXDIST가 의미하는 검색거리이며, (b)는 R의 한 변을 포함할 수 있는 거리 중에서 가장 최소인 거리를 의미한다. (a)는 두 개의 최소경계사각형 간의 마주하는 대각선상의 꼭지점을 고려하여 객체 또는 부검색공간이 존재할 수 있는 거리를 계산하기 때문에 (b)보다 검색거리가 더 큼을 알 수 있다. 따라서 대각선상의 꼭지점을 이용한 MINMAXDIST 방법은 최소근접질의 처리시 색인에서 방문되는 노드들을 줄이기 위한 정확한 검색거리 측도가 되지 못한다.



(그림 1) 검색거리의 예
(Fig. 1) example of search distance

셋째, 객체 또는 부검색공간이 차지하는 영역을 나타내는 최소경계사각형들 간의 겹침을 고려하지 않았다. 그러나 R-트리 유형의 색인방법은 최소경계사각형들 간에 겹침이 많이 발생하므로 이를 고려한 최소근접질의 처리 방법이 필요하다.

넷째, 최소근접질의에 따른 처리 비용을 최소화하기 위한 공간색인구조의 구현 및 성능평가가 이루어지지 않았다.

3. 최적탐색거리를 이용한 최소근접질의

본 장에서는 최소근접질의 처리를 위한 기본개념과 N차원 검색공간에서 기존의 측정 방법이 가지고 있는 문제점을 해결한 최적탐색거리의 특성을 정리한다.

3.1 기본 개념

최소근접질의의 처리 비용은 질의처리시 검색되는 노드의 수에 크게 좌우된다. 검색되는 노드의 수를 최소화하기 위해서는 질의 기준의 유형에 관계없이 방문되는 노드를 정확히 선정할 수 있는 오차 없는 검색거리 측도가 필요하다. 이를 위해 확장최소거리(extended MINimum DISTance : XMINDIST)와 최적탐색거리(the Optimized MiniMum value of all DISTances : OMMDIST)가 제안되었다[13]. 그리고 [13]에서는 최소근접질의 처리시 XMINDIST와 OMMDIST의 이용에 따른 질의처리 결과의 정확성을 2차원 검색공간에서 정리하고 그 특성을 보였다.

정의된 XMINDIST는 N차원 검색공간에서 최소경계사각형 M에 포함되어 있는 부검색공간들 중에서 질의 기준 Q에 가장 근접하고 있는 객체 또는 부검색공간을 결정하기 위한 거리이다. 그리고 OMMDIST는 질의 기준 Q로부터 최소경계사각형 M을 구성하는 임의의 N-1차원을 포함할 수 있는 거리들 중 최소거리로 계산된다. 예를 들면, 2차원 검색공간에서 OMMDIST는 Q로부터 M의 임의의 한 변을 포함할 수 있는 거리들 중 최소거리가 된다.

최소근접질의 알고리즘에서는 R-트리를 검색하는 동안 방문할 필요가 없는 노드들을 방문대상에서 제외하기 위해 XMINDIST와 OMMDIST를 조합하여 다음과 같은 전략을 사용한다.

i) 질의 기준 Q로부터 최소경계사각형 M'까지의 OMMDIST(Q,M')보다 XMINDIST(Q,M)가 더 큰 값을 갖는 최소경계사각형 M이 존재하면, M은 최근접객체를 포함하지 않기 때문에 M에 해당하는 노드는 검색대상에서 제외한다.

ii) 질의 기준 Q로부터 객체 O까지의 거리가 최소경계사각형 M에 대한 OMMDIST(Q,M)보다 크다면, 객체 O를 최근접객체 대상에서 제외한다.

iii) 질의 기준 Q로부터 객체 O까지의 거리보다 더 큰 XMINDIST(Q,M)를 갖는 모든 최소경계사각형 M은 검색대상에서 제외한다.

3.2 최적탐색거리의 특성

다음의 정의는 다차원 검색공간을 구성하는 임의의 도메인에서 질의 기준 Q와 최소경계사각형 M이 차지하는 범위간의 공간관계를 표현하기 위한 것이며, [13]에서 정의된 공간관계를 다음의 정리들을 위해 수정한 것이다.

즉, 2차원 검색공간에서는 Q를 구성하는 변과 M을 구성하는 변과의 공간관계를 의미하며, 3차원 검색공간에서는 Q와 M을 각각 구성하는 어느 하나의 축에 직각인 평면 사각형간의 공간관계를 의미한다. 그리고 N차원에서는 임의의 한 차원을 중심으로 초월평면과의 공간관계를 의미한다.

최적탐색거리의 특성 및 기존의 최대검색거리와의 차이는 다음의 네 가지 공간관계를 이용하여 평가된다.

【정의 1】 N차원 검색공간을 구성하는 임의의 도메인 D_i에서 질의 기준 Q와 최소경계사각형 M이 차지하는 폐쇄 경계 구간을 $I_i(Q)$와 $I_i(M)$이라 하자. $I_i(Q) = [Q_{Li}, Q_{Ui}]$이고 $I_i(M) = [M_{Li}, M_{Ui}]$일 때, $I_i(Q)$와 $I_i(M)$의 공간관계를 다음과 같이 IR_{1_i}, IR_{2_i}, IR_{3_i}, IR_{4_i}로 정의한다($1 \le i \le N$).

(1) $I_i(Q)$와 $I_i(M)$이 다음을 만족하면, $I_i(Q)$와 $I_i(M)$의 공간관계를 IR_{1_i}이라고 한다.

$$Q_{Ui} \leq M_{Li} \text{이거나 } M_{Ui} \leq Q_{Li}$$

(2) $I_i(Q)$와 $I_i(M)$이 다음의 두 가지 경우중 하나를 만족하면, $I_i(Q)$와 $I_i(M)$의 공간관계를 IR_{2_i}이라고 한다.

1) $Q_{Li} < M_{Li}$이고 $Q_{Ui} < M_{Ui}$
2) $M_{Li} < Q_{Li}$이고 $Q_{Ui} > M_{Ui}$

(3) $I_i(Q)$와 $I_i(M)$이 다음을 만족하면, $I_i(Q)$와 $I_i(M)$의 공간관계를 IR_{3_i}이라고 한다.

$$Q_{Li} \leq M_{Li} \text{이고 } M_{Ui} \leq Q_{Ui}$$

(4) $I_i(Q)$와 $I_i(M)$이 다음의 세 가지 경우중 하나를 만족하면, $I_i(Q)$와 $I_i(M)$의 공간관계를 IR_{4_i}이라고 한다.

1) $Q_{Li} = M_{Li}$이고 $Q_{Ui} < M_{Ui}$
2) $M_{Li} < Q_{Li}$이고 $Q_{Ui} < M_{Ui}$
3) $M_{Li} < Q_{Li}$이고 $Q_{Ui} = M_{Ui}$ □

다음의 보조정리에서는 N차원 검색공간에서 Q와 M을 이루는 폐쇄 경계 구간의 관계가 M을 기준으로 $Q_{Li} \geq M_{Ui}$ 또는 $\frac{(Q_{Li}+Q_{Ui})}{2} \geq \frac{(M_{Li}+M_{Ui})}{2}$ ($i = 1, \cdots, N$)인 경우를 고려한다. 그러나 기술될 보조정리는 다른 모든 관계로 확장하여 동일하게 증명될 수 있다.

【보조정리 1】 N차원 검색공간에서 질의 기준 Q로 부터 최소경계사각형 M까지의 최적탐색거리 $OMMDIST$는 M이 차지하고 있는 임의의 폐쇄 경계 구간으로 구성되는 $N-1$차원을 적어도 하나는 포함 하는 거리이다.

(증명) N차원 검색공간에서 질의 기준 Q와 최소경계 사각형 M은 다음과 같이 폐쇄 경계 구간들의 카테시언 곱으로 기술할 수 있다.

$$Q = I_1(Q) \times I_2(Q) \times \cdots \times I_N(Q)$$
$$M = I_1(M) \times I_2(M) \times \cdots \times I_N(M)$$

여기에서

$I_i(Q) = [Q_{Li}, Q_{Ui}]$, $I_i(M) = [M_{Li}, M_{Ui}]$

Q_{Li}, Q_{Ui} : 도메인 D_i에서 Q가 차지하는 영역의 시작과 끝,

M_{Li}, M_{Ui} : 도메인 D_i에서 M이 차지하는 영역의 시작과 끝.

$$Q_{Li} \leq Q_{Ui},\ M_{Li} \leq M_{Ui},\ 1 \leq i \leq N.$$

증명에서는 $OMMDIST$의 연산시 폐쇄 경계 구간으로 구성된 M의 $N-1$차원이 포함됨을 보이면 된다.

경우 1. 질의 기준 Q와 최소경계사각형 M이 겹치지 않는 경우

Q와 M를 이루는 폐쇄 경계 구간의 관계가 $Q_{Li} \geq M_{Ui}$ 또는 $\frac{(Q_{Li}+Q_{Ui})}{2} \geq \frac{(M_{Li}+M_{Ui})}{2}$ ($i = 1, \cdots, N$)이므로 $OMMDIST$는 다음과 같다. 그리고 $qr_i = (Q_{L1}, Q_{L2}, \cdots, Q_{LN})$ or $(Q_{U1}, Q_{U2}, \cdots, Q_{UN})$, $Qr_k = (Q_{L1}, Q_{L2}, \cdots, Q_{LN})$, $mr_i = (M_{U1}, M_{U2}, \cdots, M_{UN})$, $Mr_k = (M_{L1}, M_{L2}, \cdots, M_{LN})$이다.

$$OMMDIST = \min_{1 \leq i \leq N} \{ (|Q_m - M_{Ui}|^2 + \sum_{\substack{i \neq k \\ 1 \leq k \leq N}} |Q_{Lk} - M_{Lk}|^2),$$
$$(|Q_m - M_{Li}|^2 + \sum_{\substack{i \neq k \\ 1 \leq k \leq N}} |Q_{Lk} - M_{Uk}|^2) \}$$

여기에서 $Q_m = Q_{Li}$ or Q_{Ui}이다.

그런데 $|Q_{Lm} - M_{Lm}|^2 = \{ |Q_{Lm} - M_{Um}| + |M_{Um} - M_{Lm}| \}^2$ ($m = i, k$)이고, $|Q_m - M_{Li}|^2 = \{ |Q_m - M_{Ui}| + |M_{Ui} - M_{Li}| \}^2$이기 때문에 $OMMDIST$는 다음과 같이 기술할 수 있다.

$$OMMDIST = \min_{1 \leq i \leq N} \{ (|Q_m - M_{Ui}|^2 + \sum_{\substack{i \neq k \\ 1 \leq k \leq N}} \{|Q_{Lk} - M_{Uk}| + |M_{Uk} - M_{Lk}|\}^2),$$
$$(\{|Q_m - M_{Ui}| + |M_{Ui} - M_{Li}|\}^2 + \sum_{\substack{i \neq k \\ 1 \leq k \leq N}} |Q_{Lk} - M_{Uk}|^2) \}$$
$$= \min_{1 \leq i \leq N} \{ (|Q_m - M_{Ui}|^2 +$$

$$\sum_{\substack{i\neq k\\ 1\le k\le N}} \{|Q_{Lk}-M_{Uk}|+I_k(M)\}^2),$$

$$(\{|Q_m-M_{Ui}|+I_i(M)\}^2+$$

$$\sum_{\substack{i\neq k\\ 1\le k\le N}} |Q_{Lk}-M_{Uk}|^2)\}$$

따라서 2차원에서는 M의 한 변을 의미하는 폐쇄 경계 구간 I_1 또는 I_2가 $OMMDIST$에 포함된다. 3차원에서는 M의 한 면을 의미하는 폐쇄 경계 구간의 카테시언 곱 $I_1 \times I_2$, $I_1 \times I_3$, $I_2 \times I_3$중에서 하나가 $OMMDIST$에 포함된다. 그리고 i차원에서는 M의 $i-1$차원을 의미하는 폐쇄 경계 구간의 카테시언 곱 $I_1 \times I_3 \times \cdots \times I_i$, $I_1 \times I_2 \times I_4 \times \cdots \times I_i$, $I_2 \times I_3 \times \cdots \times I_i$, $\cdots$, $I_1 \times I_2 \times \cdots \times I_{i-1}$중에서 하나가 $OMMDIST$에 포함된다.

경우 2. 질의 기준 Q와 최소경계사각형 M이 겹치는 경우

Q와 M를 이루는 폐쇄 경계 구간의 관계가 $Q_{Li} \ge M_{Ui}$ 또는 $\frac{(Q_{Li}+Q_{Ui})}{2} \ge \frac{(M_{Li}+M_{Ui})}{2}$ ($i=1, \cdots, N$)이므로 $OMMDIST$는 다음과 같다. 그리고 $qr_i=(Q_{L1}, Q_{L2}, \cdots, Q_{LN})$, $Qr_k=(Q_{U1}, Q_{U2}, \cdots, Q_{UN})$, $q_i=(Q_{L1}, Q_{L2}, \cdots, Q_{LN})$, $Q_k=(Q_{L1}, Q_{L2}, \cdots, Q_{LN})$, $mr_i=(M_{U1}, M_{U2}, \cdots, M_{UN})$, $Mr_k=(M_{L1}, M_{L2}, \cdots, M_{LN})$이다.

$$OMMDIST = \min_{1\le i\le N}\{(|Q_{Li}-M_{Ui}|^2+$$

$$\sum_{\substack{i\neq k\\ 1\le k\le N}} |Q_{Uk}-M_{Lk}|^2),$$

$$(|Q_{Li}-M_{Li}|^2+\sum_{\substack{i\neq k\\ 1\le k\le N}} |Q_{Uk}-M_{Uk}|^2),$$

$$(|Q_{Li}-M_{Li}|^2+\sum_{\substack{i\neq k\\ 1\le k\le N}} |Q_{Lk}-M_{Uk}|^2)\}$$

그런데 $|Q_{Ui}-M_{Li}|^2=\{|Q_{Ui}-M_{Ui}|+|M_{Ui}-M_{Li}|\}^2$이고, $OMMDIST > |Q_{Li}-M_{Li}|$이다. 또한 $OMMDIST$는 (M_{Li}, M_{Uk})을 포함하는 거리이기 때문에 다음과 같이 기술할 수 있다.

$$OMMDIST = \min_{1\le i\le N}\{(|Q_{Li}-M_{Ui}|^2+$$

$$\sum_{\substack{i\neq k\\ 1\le k\le N}} \{|Q_{Uk}-M_{Uk}|+|M_{Uk}-M_{Lk}|\}^2),$$

$$(|Q_{Li}-M_{Li}|^2+\sum_{\substack{i\neq k\\ 1\le k\le N}} |Q_{Uk}-M_{Uk}|^2),$$

$$(|Q_{Li}-M_{Li}|^2+\sum_{\substack{i\neq k\\ 1\le k\le N}} |Q_{Lk}-M_{Uk}|^2)\}$$

$$= \min_{1\le i\le N}\{(|Q_{Li}-M_{Ui}|^2+$$

$$\sum_{\substack{i\neq k\\ 1\le k\le N}} \{|Q_{Uk}-M_{Uk}|+I_k(M)\}^2),$$

$$(|Q_{Li}-M_{Li}|^2+\sum_{\substack{i\neq k\\ 1\le k\le N}} |Q_{Uk}-M_{Uk}|^2),$$

$$(|Q_{Li}-M_{Li}|^2+\sum_{\substack{i\neq k\\ 1\le k\le N}} |Q_{Lk}-M_{Uk}|^2)\}$$

따라서 2차원에서는 M의 한 변을 의미하는 폐쇄 경계 구간 I_1 또는 I_2가 $OMMDIST$에 포함된다. 3차원에서는 M의 한 면을 의미하는 폐쇄 경계 구간의 카테시언 곱 $I_1 \times I_2$, $I_1 \times I_3$, $I_2 \times I_3$중에서 하나가 $OMMDIST$에 포함된다. 그리고 i차원에서는 M의 $i-1$차원을 의미하는 폐쇄 경계 구간의 카테시언 곱 $I_1 \times I_3 \times \cdots \times I_i$, $I_1 \times I_2 \times I_4 \times \cdots \times I_i$, $I_2 \times I_3 \times \cdots \times I_i$, $\cdots$, $I_1 \times I_2 \times \cdots \times I_{i-1}$ 중에서 하나가 $OMMDIST$에 포함된다.

이상과 같이 $OMMDIST$는 M을 구성하는 임의의 한 폐쇄 경계 구간 또는 폐쇄 경계 구간의 카테시언 곱으로 나타내는 $N-1$차원을 반드시 포함하는 거리이므로, N차원 검색공간에서 질의 기준 Q로부터 최소경계사각형 M까지의 $OMMDIST$는 M을 구성하는 임의의 한 폐쇄 경계 구간이 적어도 하나는 포함되는 거리임이 성립한다. □

증명에서 사용된 기호 qr_i, Qr_k, q_i, Q_k, mr_i, M_k 들의 의미는 다음과 같다. 질의기준 Q와 최소경계사각형 M이 겹치지 않는 경우에 qr_i, Qr_k는 M의 중심에서 가장 가까운 Q의 꼭지점이나, Q의 평면들 중 각 차원 축과 직교하는 두 개의 평면들 중 가장 가까운 점과 가장 먼 점이 된다. Q와 M이 겹치는 경우에 qr_i, Qr_k, q_i, Q_k

는 M의 중심에서 가장 가까운 Q의 복지점 또는 Q의 평면들 중 각 차원 축과 직교하는 두 개의 평면들 중 가장 먼 점이 된다. 그리고 mr_k, M_k는 Q의 중심에서 가장 가까운 M의 평면들 중 각 차원 축과 직교하는 두 개의 평면들 중 가장 먼 점이 된다.

【보조정리 2】 N차원 검색공간에 있는 최소경계사각형 M을 구성하는 임의의 한 $N-1$ 차원에서 그 $N-1$ 차원에 접해 있는 객체 또는 부검색공간이 반드시 하나는 존재한다.

(증명) N차원 검색공간에서 M에 포함되어 있는 m개의 객체 또는 부검색공간들을 S라 하자.

$S = \{s_1, s_2, \cdots, s_m\}$ (s_i : i번째 객체 또는 부검색공간의 영역)

그러면 M은 다음과 같이 나타낼 수 있다.

$$M = [S_L, S_U]_1 \times [S_L, S_U]_2 \times \cdots \times [S_L, S_U]_N$$

여기에서

$S_{Lj} = \min\{s_i(D_j)\}$, $S_{Uj} = \max\{s_i(D_j)\}$

$s_i(D_j)$: s_i를 도메인 D_j에 사상한 값들이다 $(1 \le i \le m,\ 1 \le j \le N)$.

S가 M의 임의의 폐쇄 경계 구간으로 구성된 어느 하나의 $N-1$ 차원에 접하지 않는다고 가정하자. 그러면 M에 대한 폐쇄 경계 구간의 값 S_{Lj}, S_{Uj}중에서 적어도 하나는 다음의 1) 또는 2)를 만족하게 된다.

1) $S_{Lj} < \min\{s_i(D_j)\}$

2) $S_{Uj} > \max\{s_i(D_j)\}$

위의 어떤 경우에도 최소경계사각형 정의에 모순이 되기 때문에 기존의 M보다 더 작은 최소경계사각형이 생성되게 된다. 따라서 최소경계사각형 M의 임의의 한 $N-1$ 차원에는 반드시 접해 있는 객체 또는 부검색공간이 존재한다. □

【정리 1】 N차원 검색공간에서 질의 기준 Q로부터 최소경계사각형 M까지의 최적탐색거리 $OMMDIST$는 M이 차지하고 있는 임의의 폐쇄 경계 구간으로 구성되는 $N-1$ 차원을 적어도 하나는 포함하는 거리이며, 그 $N-1$ 차원에 접해 있는 객체 또는 부검색공간이 반드시 하나는 존재한다.

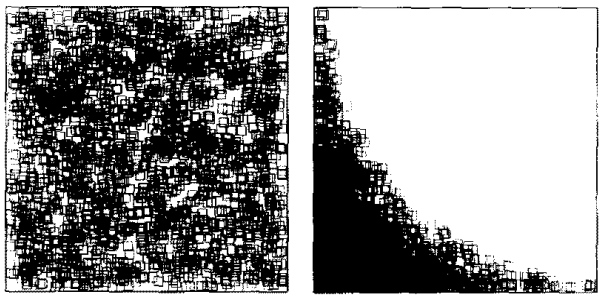
(증명) 보조정리 1에 의해 Q로부터 M까지의 $OMMDIST$가 M을 구성하는 $N-1$ 차원을 적어도 하나는 포함하는 거리임이 증명되며, 보조정리 2에 의해 M의 임의의 한 $N-1$차원에 객체 또는 부검색공간이 반드시 접해 있음이 증명된다. 따라서 위의 정리가 성립함이 증명된다. □

정리 1은 본 논문에서 제안한 검색거리 측도의 기본 개념 및 특성이다. N차원 검색공간에서 최소근접질의 처리시 질의 기준 Q로부터 최소경계사각형을 구성하는 임의의 한 N-1차원을 반드시 포함하는 거리인 OMMDIST을 적용하므로써 검색될 노드를 정확히 선택할 수 있다. 따라서 색인에서 방문되는 전체 노드의 수가 최소화되기 때문에 질의처리에 따른 비용을 최적화한다.

최적탐색거리를 이용하면 검색공간에 존재하는 객체들에 대한 검색범위를 가장 작게 유지하므로 트리에서 검색되는 노드의 수를 줄인다. 따라서 최소근접질의 처리시 보조기억장치의 접근횟수를 최소화할 수 있다.

4. 실험을 통한 성능 평가

본 장에서는 최적탐색거리 OMMDIST의 성능을 질의 기준으로부터 가장 가까운 검색대상 노드들을 찾는데 소요되는 디스크 접근 횟수와 질의 처리 시간에 따라 평가한다. 실험에서는 이차비용 분할 알고리즘을 사용한 R-트리[5]에 최적탐색거리 OMMDIST와 최대검색거리 MINMAXDIST에 의한 최소근접질의 처리 알고리즘[13]을 적용한 후 이에 따른 실험결과에 의해 그 성능을 비교한다.



(a) 균일분포 (b) 지수분포

(그림 2) 객체 분포
(Fig. 2) object distributions

실험에서는 이차원 검색공간에서 중복되지 않은 20,000개의 사각형 객체를 사용하였으며, 논문에서 제안한 검색거리 측도의 성능을 명확히 보이기 위해 동일한 비율 및 80바이트의 고정된 크기를 갖는 사각형 객체를 가정하였다. 또한 객체들의 면적은 전체 검색공간의 0.1%, 0.001%인 경우를 고려하였다. 이는 색인에 삽입되는 대부분의 이는 정도 겹치는 경우(0.1%)와 서로 겹치지 않는 경우(0.001%)에 대해서 검색거리 측도의 특성을 보이기 위해서이다.

하나의 도메인의 범위가 [0,1)이라할 때, 실험에서 사용된 객체의 분포는 다음과 같다(그림 2 참조).

- 균일분포(uniform distribution) : 사각형 객체들의 중심점이 중복되지 않고 균일하게 분포
- 지수분포(exponential distribution) : 평균 0.5인 지수 함수 분포

균일분포는 객체의 분포가 이상적인 상태에서의 최소근접질의의 처리성능을 규명하기 위한 것이다. 그리고 지수분포는 객체들이 검색공간의 일부분에 집중되어 발생하는 경우에 최소근접질의의 처리성능을 규명하기 위한 것이다. 디스크에 저장된 R-트리의 노드와 버켓은 동일한 크기를 가지며, 모든 노드와 버켓의 접근시간은 동일하다고 가정한다.

- 노드 및 버켓의 헤더 크기 : 16바이트
- MBR 크기 : 16바이트
- 주소지시자 크기 : 4바이트

색인 구축을 위한 노드의 분기율과 버켓용량은 다음과 같이 계산된다. 그리고 실험에서는 100개의 최소근접질의를 균일하게 발생시켜 디스크 접근 횟수와 질의 처리시간에 따른 성능을 알아보았으며, 질의 크기는 전체 검색공간의 0.01%, 1.0%으로 하였다.

$$\text{노드 분기율} = \frac{\text{노드 크기} - \text{노드 헤더 크기}}{\text{영역 좌표 크기} + \text{주소지시자 크기}}$$

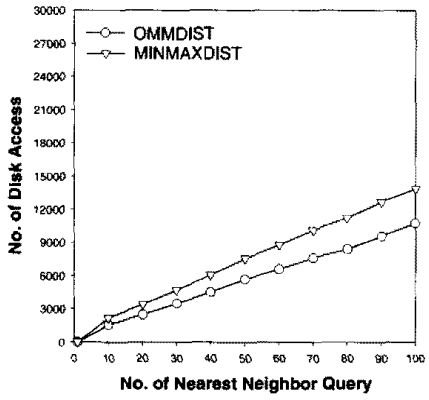
$$\text{버켓용량} = \frac{\text{버켓 크기} - \text{버켓 헤더 크기}}{\text{객체 크기}}$$

그림 3과 그림 4는 균일분포와 지수분포에서 객체의 크기가 0.1%일 때 질의 수의 증가에 따른 디스크 접근 횟수를 나타낸 것이다. 그림에서는 질의의 갯수가 많아질수록 디스크 접근 횟수가 선형적인 증가를 보이며, OMMDIST와 MINMAXDIST의 성능 차이가 더 커짐을 보인다. 그리고 객체들이 검색공간의 일부분에 집중되어 발생되는 분포에서는 그 차이가 더 커짐을 알 수 있다.

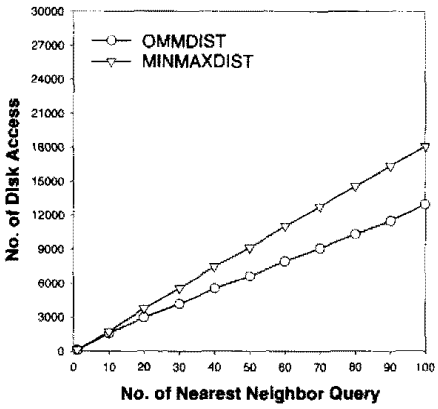
반복된 실험결과에 의하면 OMMDIST와 MINMAXDIST에 따른 디스크 접근 횟수는 질의의 크기가 크고, 객체들이 검색공간의 한 영역에 집중적으로 발생하는 분포일수록 디스크 접근 횟수가 많아지는 경향을 보인다.

그러나 OMMDIST를 이용한 처리 방법은 질의 크기나 객체의 분포에 상관없이 MINMAXDIST에 의한 처리 방법보다 항상 낮은 디스크 접근 횟수를 나타낸다. 그리고 각각의 분포에서 질의의 크기가 1.0%일 때에는 OMMDIST와 MINMAXDIST의 디스크 접근 횟수의 차가 더 커짐을 알 수 있었다.

이는 색인에서 검색대상이 되는 노드의 선택시 OMMDIST에 의한 검색거리가 MINMAXDIST에 의한 검색거리 보다 더 작기 때문이다. 따라서 검색대상에

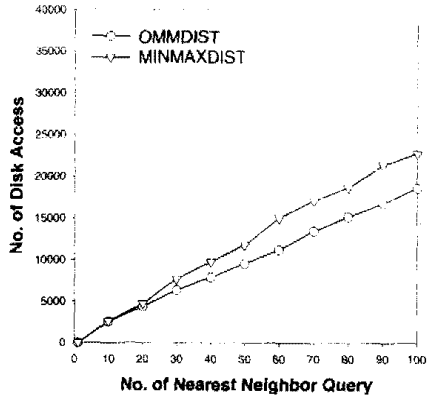


(a) 질의 크기 = 0.01%

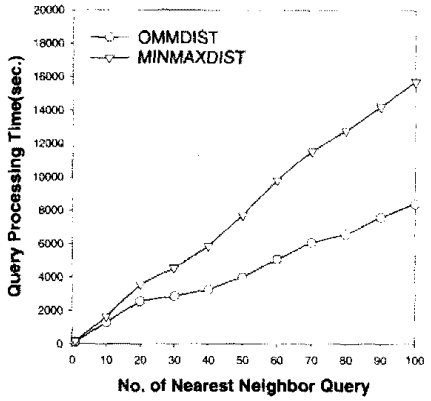


(b) 질의 크기 = 1.0%

(그림 3) 균일분포에서 디스크 접근 횟수(버켓용량 =12)
(Fig. 3) number of disk access in uniform distribution

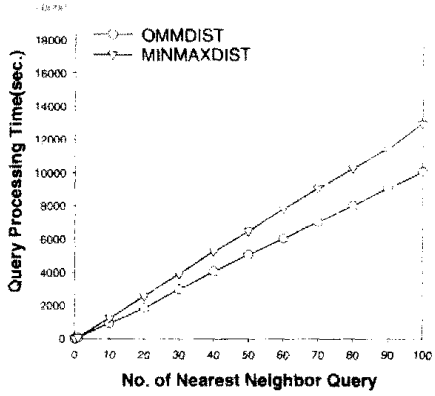


(a) 질의 크기 = 0.01%

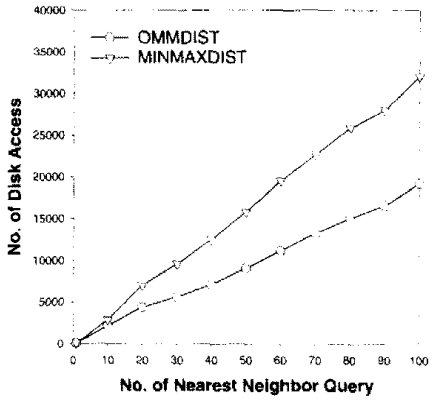


(b) 지수분포(버켓용량=50)

(그림 4) 지수분표에서 디스크 접근 횟수(버켓용량=12)
(Fig. 4) number of disk access in exponential distribution



(a) 균일분포(버켓용량=50)



(b) 질의 크기 = 1.0%

(그림 5) 최소근접질의 처리 시간
(Fig. 5) processing time of nearest neighbor queries

서 제외되는 노드들이 MINMAXDIST에 의한 방법보다 OMMDIST에 의한 방법이 더 많기 때문에 결과적으로 디스크 접근 횟수가 더 적어지게 된다.

이러한 결과는 OMMDIST를 이용하여 검색거리를 정확히 측정함으로써 검색되는 노드 및 버켓의 수를 줄이기 때문이다.

그림 5는 최소근접질의에 따른 질의 처리 시간을 분석하기 위해 객체의 크기가 0.001%이고 질의 크기가 1.0%일 때 100개의 최소근접질의를 처리하는데 따른 누적된 질의 처리 시간을 나타낸 것이다.

그림에서는 객체들이 검색공간에서 거의 겹치지 않고 한 영역에 집중적으로 발생하는 지수분포의 경우 OMMDIST를 이용한 처리 방법과 MINMAXDIST를 이용한 처리 방법과의 질의 처리 시간의 차가 더 커짐을 보여준다. 이러한 이유는 색인에서 방문대상이 되는 노드들의 선정시 MINMAXDIST에 의한 방법이 OMMDIST에 의한 방법보다 더 많은 노드들을 선택하기 때문이다. 그러므로 질의의 수가 많아질수록 검색되는 노드의 수가 가중되어 이에 따른 질의 처리 시간이 더욱 커지게 된다.

반복된 실험결과에 의하면 객체들의 모든 분포에 있어서 대부분의 객체가 겹치는 경우인 1.0%일 때에는 OMMDIST를 이용한 방법과 MINMAXDIST를 이용한 방법의 질의 처리 시간은 비슷한 경향을 보임을 알 수 있었다.

4. 결 론

최소근접질의 처리시 색인에서 검색되는 노드들을 정확히 선정할 수 있는 최적탐색거리의 이용은 질의처리 비용을 최소화할 수 있다. 본 논문에서는 N차원 검색공간에서 최적탐색거리의 특성을 정리하였으며, 실험을 통하여 기존의 방법과 질의처리 성능을 비교 분

입하였다. 증명을 통하여 N차원 검색공간에서도 최적탐색거리는 색인에서 방문된 노드들을 정확히 선택할 수 있음을 보였다. 따라서 최적탐색거리의 이용은 검색되는 노드의 수를 최소화하기 때문에 최소근접질의에 따른 처리 비용을 최적화한다.

실험에서는 최적탐색거리의 성능을 디스크 접근 횟수와 질의처리시간에 따라 비교 평가하였다. 실험 데이타로는 균일분포와 지수분포를 이루는 사각형 객체들을 이용하였다. 실험결과에 의하면, 최적탐색거리를 이용한 최소근접질의의 처리는 객체의 분포형태, 객체의 크기 그리고 버켓의 용량에 관계없이 항상 낮은 디스크 접근 횟수를 보였다. 특히 질의의 크기가 클 때에는 최적탐색거리를 이용한 최소근접질의의 처리가 최대검색거리를 이용한 처리에 비해 디스크 접근 횟수가 더욱 작아 진다. 또한 질의의 수가 많아질수록 질의처리 시간이 상대적으로 더 적어짐을 알 수 있었다.

참 고 문 헌

[1] N.Beckmann, H.Kriegel, R.Schneider and B.Seeger, "The R*-tree : a Efficient and Robust Access Method for Points and Rectangles," Proc. ACM SIGMOD Int. Conf. on Management of Data, pp.322-331, 1990.

[2] T.Brinkhoff, H.P.Kriegel and R.Schneider, "Comparison of Approximation of Complex Objects Used for Approximation-based Query Processing in Spatial Database Systems," Proc of the 9th Int. Conf. on Data Engineering, pp.40-49, 1993.

[3] T.Brinkhoff, H.P.Kriegel, R.Schneider and B.Seeger, "Multi-Step Processing of Spatial Joins," Proc. ACM SIGMOD Int. Conf. on Management of Data, 1994.

[4] O.Guenther and A.Buchmann, "Research Issues in Spatial Databases," ACM SIGMOD Record, Vol.19, No.4, pp.61-68, 1990.

[5] A.Guttman, "R-tree: a Dynamic Index Structure for Spatial Searching," Proc. ACM SIGMOD Int. Conf. on Management of Data, pp.599-608, 1984.

[6] R.H.Güting, "An Introduction to Spatial Database Systems," VLDB Journal, No.3, pp.357-399, Aug. 1994.

[7] E.G.Hoel and H.Samet, "Efficient Processing of Spatial Queries in Line Segment Databases," Proc. of the 2nd Sym. on Large Spatial Databases, pp.237-256, 1991.

[8] N.Katayama and S.Satoh, "The SR-Tree: An Index Structure for HighDimensional Nearest Neighbor Queries," Proc. ACM SIGMOD Int. Conf. on Management of Data, pp.369-380, 1997.

[9] H.P.Kriegel, H.Horn and M.Schiwietz, "The Performance of Object Decomposition Techniques for Spatial Query Processing," Proc. of the 2nd Sym. on Large Spatial Databases, pp.257-276, 1991.

[10] C.B.Mederios and F.Pires, "Databases for GIS," Proc. ACM SIGMOD Int. Conf. on Management of Data, Vol.23, No.1, pp.107-115, 1994.

[11] D.Papadias, Y.Theodoridis, T.Sellis, and M.J. Egenhofer, "Topologocal Relations in the World of Minimum Bounding Rectangles: A Study with R-trees," Proc. ACM SIGMOD Int. Conf. on Management of Data, pp.92-103, 1995.

[12] N.Roussopoulos, S.Kelley and F.Vincent, "Nearest Neighbor Queries," Proc. ACM SIGMOD Int. Conf. on Management of Data, pp.71-79, 1995.

[13] 선휘준, 황부현, 류근호, "최적탐색거리를 이용한 최근접질의의 처리 방법", 한국정보처리학회논문지, 제4권 9호, pp.2173-2184, 1997. 9.

[14] 이동만, 이용주, 정진완, "R-트리를 이용한 최근접 근질의 처리에 관한 연구", 한국정보과학회 추계학술발표논문집, 제23권, 2호, pp.35-38, 1996.

선 휘 준

e-mail : hjseon@tiger.seonam.ac.kr

1988년 목포대학교 전산통계학과 졸업

1990년 전남대학교 대학원 전산통계학과(이학석사)

1998년 전남대학교 대학원 전산통계학과(이학박사)

1997년~현재 서남대학교 전산정보학과 전임강사

관심분야 : 공간 데이타베이스, 공간자료구조, 시공간데이타베이스, 지리정보시스템

김 홍 기

e mail : hkkim@dongshinu.ac.kr

1984년 전남대학교 계산통계학과 (이학사)

1986년 전남대학교 대학원 계산통계학과(이학석사)

1996년 전남대학교 대학원 전산통계학과(이학박사)

1991년 ~ 현재 동신대학교 전산통계학과 조교수

관심분야 : 공간데이타구조, 공간데이타베이스, 컴퓨터 그래픽스